# Varieties

*Howard Straubing*[1]*and Pascal Weil*[2]

[1]Computer Science Department
Boston College
Chestnut Hill, Massachusetts 02467, USA
email: straubin@cs.bc.edu

[2]LaBRI
Université de Bordeaux and CNRS
351 cours de la Libération, 33405 Talence Cedex, France
email: pascal.weil@labri.fr

This chapter is devoted to the theory of varieties, which provides an important tool, based in universal algebra, for the classification of regular languages. In the introductory section, we present a number of examples that illustrate and motivate the fundamental concepts. We do this for the most part without proofs, and often without precise definitions, leaving these to the formal development of the theory that begins in Section 2. Our presentation of the theory draws heavily on the very recent work of Gehrke, Grigorieff and Pin [22] on the equational theory of lattices of regular languages. In the subsequent sections we consider in more detail aspects of varieties that were only briefly evoked in the introduction: Decidability, operations on languages, and characterizations in formal logic.

## 1 Motivation and examples

We refer the readers to Chapter 1, and specifically to Sections 4.2 and 4.3 of that chapter, for the notion of a language recognized by a morphism into a finite monoid, and for the definition of the syntactic monoid $\mathrm{Synt}(L)$ of a language $L$.

### 1.1 Idempotent and commutative monoids

When one begins the study of abstract algebra, groups are usually encountered before semigroups and monoids. The simplest example of a monoid that is not a group is the set $\{0, 1\}$ with the usual multiplication. We denote this monoid $U_1$.

*Work partially supported by NSF Grant CCF-0915065
†Work partially supported by ANR Grant FREC 2010 BLAN 0202

What are the regular languages recognized by $U_1$? If $A$ is a finite alphabet and $\varphi\colon A^* \to U_1$ is a morphism, then any language $L \subseteq A^*$ recognized by $\varphi$—that is, any set of the form $\varphi^{-1}(X)$ where $X \subseteq U_1$—has either the form $B^*$ or $A^* \setminus B^*$, where $B \subseteq A$. In particular, membership of a word $w$ in $L$ depends only on the set $\alpha(w)$ of letters occurring in $w$ (see Example 4.9 in Chapter 1).

The property 'membership of $w$ in $L$ depends only on $\alpha(w)$' is preserved under union and complement, and thus defines a boolean algebra of regular languages. Of course, not every language in this boolean algebra is recognized by $U_1$; for example, we could take $L = a^* \cup b^*$. However, it follows from basic properties of the syntactic monoid that this boolean algebra consists of precisely the languages recognized by finite direct products of copies of $U_1$.

We have thus characterized a syntactic property of regular languages in terms of an algebraic property of its syntactic monoid. The family of finite monoids that divide a direct product of a finite number of copies of $U_1$ is itself closed under finite direct products and division. Such a family of finite monoids is called a *pseudovariety*. This particular pseudovariety is often denoted $\mathbf{J}_1$ in the literature[1].

**1.1.1 Decidability and equational description** Thus if we want to decide whether a given language $L \subseteq A^*$ has this syntactic property, we can compute $\mathrm{Synt}(L)$ and try to determine whether $\mathrm{Synt}(L) \in \mathbf{J}_1$. But how do we do *that?* There are, after all, infinitely many monoids in $\mathbf{J}_1$. We can, however, bound the size of the search space in terms of $|A|$. It is not hard to prove that if $M$ is a finite monoid, and

$$\varphi\colon A^* \to \underbrace{M \times \cdots \times M}_{r \quad \text{times}}$$

is a morphism, then $N = \varphi(A^*)$ embeds into

$$\underbrace{M \times \cdots \times M}_{s \quad \text{times}},$$

where $s = |M|^{|A|}$. This settles, in a not very satisfactory way, the question of deciding whether $\mathrm{Synt}(L)$ is in $\mathbf{J}_1$: The resulting 'decision procedure'—check all the divisors of $U_1^{2^{|A|}}$ and see if $\mathrm{Synt}(L)$ is isomorphic to any of them!—is of course ridiculously impractical. Fortunately, there is a better approach: $U_1$ is both commutative and idempotent (*i.e.*, all its elements are idempotents).These two properties are preserved under direct products and division, and consequently shared by all members of $\mathbf{J}_1$. That is, the idempotent and commutative monoids form a pseudovariety that contains $\mathbf{J}_1$. Conversely, every idempotent and commutative finite monoid belongs to $\mathbf{J}_1$. To see this, we make note of a fact that will play a large role in this chapter: If $M$ is a finite monoid and $\varphi\colon A^* \to M$ an onto morphism, then

$$M \prec \prod_{m \in M} \mathrm{Synt}(\varphi^{-1}(m)).$$

In particular, *every pseudovariety is generated by the syntactic monoids it contains.* We now observe that if $\alpha(w_1) = \alpha(w_2)$, and if $\varphi\colon A^* \to M$ is a morphism onto an idempotent and commutative monoid, then $\varphi(w_1) = \varphi(w_2)$, since we can permute letters

[1]It is also written **Sl** because its elements are called semilattices.

and eliminate duplications in any word $w$ without changing its value under $\varphi$. Thus each $\varphi^{-1}(m)$ satisfies our syntactic property, and so by the remark just made, $M \in \mathbf{J}_1$.

We can express '$M$ is idempotent and commutative' by saying that $M$ *satisfies the identities* $xy = yx$ and $x^2 = x$. This means that these equations hold no matter how we substitute elements of $M$ for the variables $x$ and $y$. This equational characterization of $\mathbf{J}_1$ provides a much more satisfactory procedure for determining if a monoid $M$ belongs to $\mathbf{J}_1$ : If $M$ is given by its multiplication table, then we can verify the identities in time polynomial in $|M|$.

**1.1.2 Connection to logic** Before leaving this example, we note a connection with formal logic. We express properties of words over $A^*$ by sentences of first-order logic in which variables denote positions in a word. For each $a \in A$, our logic contains a unary predicate $Q_a$, where $Q_a x$ is interpreted to mean 'the letter in position $x$ is $a$'. We allow only these formulas $Q_a x$ as atomic formulas—in particular, we do not include equality as a predicate. A sentence in this logic, for example (with $A = \{a, b, c\}$)

$$\exists x \exists y \forall z (Q_a x \wedge Q_b y \wedge \neg Q_b z)$$

defines a language over $A^*$, in this case the set of all words containing both $a$ and $b$, but with no occurrence of $c$. It is easy to see that the languages definable in this logic are exactly those in which membership of a word $w$ depends only on $\alpha(w)$.

The following theorem summarizes the results of this subsection.

**Theorem 1.1.** *Let $A$ be a finite alphabet and let $L \subseteq A^*$ be a regular language. The following are equivalent.*

*(i) Membership of $w$ in $L$ depends only on the set $\alpha(w)$ of letters appearing in $w$.*
*(ii) $\mathrm{Synt}(L) \in \boldsymbol{J}_1$, that is, $\mathrm{Synt}(L)$ divides a finite direct product of copies of $U_1$.*
*(iii) $\mathrm{Synt}(L)$ satisfies the identities $xy = yx$ and $x^2 = x$.*
*(iv) $L$ is definable by a first-order sentence over the predicates $Q_a$, $a \in A$.*

## 1.2 Piecewise-testable languages

Suppose that instead of testing for occurrences of individual letters in a word, we test for occurrences of non-contiguous sequences of letters, or *subwords*. More precisely, we say that $v = a_1 \cdots a_k$, where each $a_i \in A$, is a subword of $w \in A^*$ if

$$w = w_0 a_1 w_1 \cdots a_k w_k$$

for some $w_0, \ldots, w_k \in A^*$. We also say that the empty word 1 is a subword of every word in $A^*$. The set of all words in $A^*$ that contain $v$ as a subword is thus the regular language

$$L_v = A^* a_1 A^* \cdots a_k A^*.$$

We say that a language is *piecewise-testable* if it belongs to the boolean algebra generated by the $L_v$.

**1.2.1 Decidability and equational description** It is not clear that we can effectively decide whether a given regular language is piecewise testable. For the language class

of 1.1, we were able to settle this question by in effect observing that for every finite alphabet $A$ there were only finitely many languages of the class in $A^*$. For piecewise-testable languages, this is no longer the case. It is possible, however, to obtain an algebraic characterization of the piecewise-testable languages, and this leads to a fairly efficient decision procedure. We first note two relatively easy-to-prove facts. First, the monoids $\mathrm{Synt}(L_v)$ are all *$\mathcal{J}$-trivial*: This means that if $m, m', s, t, s', t' \in \mathrm{Synt}(L_v)$ are such that $m = s'm't'$, $m' = smt$, then $m = m'$. Second, the family **J** of $\mathcal{J}$-trivial monoids forms a pseudovariety. It follows then that the syntactic monoid of every piecewise-testable language is $\mathcal{J}$-trivial. A deep theorem, due to I. Simon [43], shows that the converse is true as well: Every language recognized by a finite $\mathcal{J}$-trivial monoid is piecewise-testable.

Clearly, we can effectively determine, from the multiplication table of a finite monoid $M$, all the pairs $(m, m') \in M \times M$ such that $m' = smt$ for some $s, t \in M$, and thus determine if $M \in$ **J**. This gives us an algebraic decision procedure for piecewise-testability.

Can the pseudovariety **J** be defined by identities in the same manner as $\mathbf{J}_1$? The short answer is 'no'. This is because satisfaction of an identity $u = v$, where $u$ and $v$ are words over an alphabet $\{x, y, \ldots\}$ of variables, is preserved by *infinite* direct products as well as finite direct products and divisors. Consider now the monoids

$$M_j = \{1, m, m^2, \ldots, m^j = m^{j+1}\}.$$

Each $M_j \in$ **J**, but $\prod_{j \geqslant 1} M_j$ contains an isomorphic copy of the infinite cyclic monoid $\{1, a, a^2, \ldots\}$, which has every finite cyclic group as a quotient. Thus every identity satisfied by all the monoids in **J** is also satisfied by all the finite cyclic groups, which are not in **J**.

In spite of this, we can still obtain an equational description of **J**, provided we adopt an expanded notion of what constitutes an identity. If $s$ is an element of a finite monoid $M$, then we denote by $s^\omega$ the unique idempotent power of $s$. We will allow identities in which the operation $x \mapsto x^\omega$ is allowed to appear; these are special instances of what we will call *profinite identities.* It is not hard to see that satisfaction of these new identities is preserved under finite direct products and quotients, and thus every set of such identities defines a pseudovariety.

For example, the profinite identity

$$x^\omega = xx^\omega$$

is satisfied precisely the finite monoids that contain no nontrivial groups. This is the pseudovariety of *aperiodic monoids*, which we denote **Ap**. Similarly, the profinite identity

$$x^\omega = 1$$

defines the pseudovariety **G** of finite groups. As was the case with **J**, neither of these pseudovarieties can be defined by a set of ordinary identities.

It can be shown that the pseudovariety **J** of finite $\mathcal{J}$-trivial monoids is defined by the pair of profinite identities

$$(xy)^\omega x = (xy)^\omega$$
$$y(xy)^\omega = (xy)^\omega,$$

or, alternatively, by the pair

$$(xy)^\omega = (yx)^\omega$$
$$xx^\omega = x^\omega.$$

**1.2.2 Connection with logic** Let us supplement the first-order logic for words that we introduced earlier with atomic formulas of the form $x < y$, which is interpreted to mean 'position $x$ is strictly to the left of position $y$'. The language $L_v$, where $v = a_1 \cdots a_k$, is defined by the sentence

$$\exists x_1 \exists x_2 \cdots \exists x_k (x_1 < x_2 \wedge x_2 < x_3 \wedge \cdots \wedge x_{k-1} < x_k \wedge Q_{a_1} x_1 \wedge \cdots \wedge Q_{a_k} x_k).$$

This is a $\Sigma_1$-sentence—one in which all the quantifiers are in a single block of existential quantifiers at the start of the sentence. It follows easily that a language is piecewise-testable if and only if it is defined by a boolean combination of $\Sigma_1$-sentences.

The following theorem summarizes the results of this subsection.

**Theorem 1.2.** *Let $A$ be a finite alphabet and let $L \subseteq A^*$ be a regular language. The following are equivalent.*

*(i)* *$L$ is piecewise testable.*
*(ii)* $\mathrm{Synt}(L) \in$ ***J****, that is,* $\mathrm{Synt}(L)$ *is $\mathcal{J}$-trivial.*
*(iii)* $\mathrm{Synt}(L)$ *satisfies the identities* $(xy)^\omega = (yx)^\omega$ *and* $xx^\omega = x^\omega$.
*(iv)* $\mathrm{Synt}(L)$ *satisfies the identities* $(xy)^\omega x = y(xy)^\omega$.
*(v)* *$L$ is definable by a boolean combination of $\Sigma_1$-sentences over the predicates $<$ and $Q_a$, $a \in A$.*

## 1.3 Pseudovarieties of monoids and varieties of languages

We tentatively extract a few general principles from the preceding discussion. These will be explored at length in the subsequent sections. Given a pseudovariety **V** of finite monoids and a finite alphabet $A$, we form the family $A^*\mathcal{V}$ of all regular languages $L \subseteq A^*$ for which $\mathrm{Synt}(L) \in$ **V**. We can think of $\mathcal{V}$ itself as an operator that associates to each finite alphabet $A$ a family of regular languages over $A$. $\mathcal{V}$ is called a *variety of languages*. (We will give a very different, although equivalent definition of this term in our formal discussion in Section 2.) From our earlier observation that pseudovarieties are generated by the syntactic monoids they contain, it follows that if **V** and **W** are distinct pseudovarieties, then the associated varieties of languages $\mathcal{V}$ and $\mathcal{W}$ are also distinct. Thus *there is a one-to-one correspondence between varieties of languages and pseudovarieties of finite monoids.*

Often we are interested in the following sort of decision problem: Given a regular language $L \subseteq A^*$, does it belong to some predefined family $\mathcal{V}$ of regular languages, for example, the languages definable in some logic? If $\mathcal{V}$ forms a variety of languages, then we can answer the question if we have some effective criterion for determining if a given finite monoid belongs to the corresponding pseudovariety **V**. (The converse is true as well: if we could decide the question about membership in the variety of languages, we would be able to decide membership in **V**.)

*Pseudovarieties are precisely the families of finite monoids defined by sets of profinite identities.* For the time being this assertion—a theorem due to Reiterman— will have to remain somewhat vague, since we haven't even come close to saying what a profinite identity actually is! Such equational characterizations of pseudovarieties are frequently the source of the decision procedures discussed above.

If $\mathcal{V}$ is a variety of languages, then, as we have seen, each $A^*\mathcal{V}$ is closed under boolean operations. Observe further that if $L \in A^*\mathcal{V}$ and $v \in A^*$, then both of the *quotient* languages

$$v^{-1}L = \{w \in A^* \mid vw \in L\}$$
$$Lv^{-1} = \{w \in A^* \mid wv \in L\}$$

are in $A^*\mathcal{V}$, because any monoid recognizing $L$ also recognizes the quotients. For the same reason, if $\varphi\colon B^* \to A^*$ is a morphism, $\varphi^{-1}(L)$ is in $B^*\mathcal{V}$. An important result, due to Eilenberg, showed that these closure properties characterize varieties of languages.

**Theorem 1.3.** *Let $\mathcal{V}$ assign to each finite alphabet $A$ a family $A^*\mathcal{V}$ of regular languages in $A^*$. $\mathcal{V}$ is a variety of languages if and only if the following three conditions hold:*

*(i) Each $A^*\mathcal{V}$ is closed under boolean operations.*
*(ii) If $L \in A^*\mathcal{V}$ and $w \in A$, then $w^{-1}L \in A^*\mathcal{V}$, and $Lw^{-1} \in A^*\mathcal{V}$.*
*(iii) If $L \in A^*\mathcal{V}$ and $\varphi\colon B^* \to A^*$ is a morphism of finitely generated free monoids, then $\varphi^{-1}(L) \in B^*\mathcal{V}$.*

This theorem can be quite useful for showing, in the absence of an explicit algebraic characterization of the corresponding pseudovariety of monoids, that a combinatorially or logically defined family of languages forms a variety. We conclude from this that such an algebraic characterization in principle exists.

Although it is somewhat involved, Theorem 1.3 is quite elementary, see [18, 32]. In the next section we will revisit the definition of varieties of languages and profinite identities in a way that will permit us to prove both Theorem 1.3 and Reiterman's theorem in a single argument.

Before we proceed with this program, we briefly describe certain classes of regular languages which admit syntactic characterizations (that is: characterizations in terms of syntactic monoids and syntactic morphisms), but which are not varieties in the sense described above.

## 1.4 Extensions

Interesting classes of regular languages frequently admit characterizations in terms of their syntactic monoids and syntactic morphisms, and the theory sketched above is meant to provide a formal setting for this algebraic classification of regular languages. However, the framework is not adequate to capture all the examples of interest that arise. Here we give three examples.

Consider, first, the family $A^*\mathcal{K}_1$ of languages $L \subseteq A^*$ for which membership of $w$ in $L$ is determined by the leftmost letter of $w$. This class forms a boolean algebra closed under quotients, but is not a variety of languages. To see this, note that $a(a + b)^* \in$

$\{a,b\}^*\mathcal{K}_1$ and $c^*a(a+b+c)^* \notin \{a,b,c\}^*\mathcal{K}_1$, even though the two languages have the same syntactic monoid. Alternatively, we can reason using Theorem 1.3, and note that the second language is an inverse homomorphic image of the first, and thus $\mathcal{K}_1$ fails to be a variety of languages. More generally, we can define the family $A^*\mathcal{K}_d$ of languages $L$ for which membership of $w$ in $L$ depends only on the leftmost $\min(|w|, d)$ letters of $w$, as well as $A^*\mathcal{K} = \bigcup_{d>0} A^*\mathcal{K}_d$. All these families are closed under boolean operations and quotients, yet fail to be varieties of languages.

We obtain an example with a similar flavor if we supplement the predicate logic described earlier by atomic formulas $x \equiv_q 0$, where $q > 1$, which is interpreted to mean that position $x$ is divisible by $q$. (We assume that positions in a word are numbered, beginning with $1$ for the leftmost position.) We denote by $A^*\mathcal{QA}$ the family of languages over $A^*$ definable in this logic. Languages in $A^*\mathcal{QA}$ arise as the regular languages definable in the circuit complexity class $AC^0$ (see [10]). Each $A^*\mathcal{QA}$ is a boolean algebra closed under quotients, however $\mathcal{QA}$ is not a variety of languages: To see this, consider the morphism $\{a,b\}^* \to \{a\}^*$ that maps $a$ to $a$ and $b$ to the empty string. The set $\{a^{2n} \mid n \geqslant 0\}$ is in $\{a\}^*\mathcal{QA}$, as it is defined by by the sentence

$$\forall x(\forall y(y \leqslant x) \to x \equiv_2 0).$$

However the inverse image of this language under the morphism is the set of strings over $\{a,b\}$ with an even number of occurrences of $a$, and it is possible to prove by model-theoretic means that this language is not definable in our logic.

Finally, consider the family $A^*\mathcal{J}^+$ of languages definable by $\Sigma_1$-sentences over the predicates $<$ and $Q_a$ with $a \in A$ (in contrast to the languages definable by boolean combinations of $\Sigma_1$-sentences, which we considered earlier). It is easy to see that if $L \in A^*\mathcal{J}^+$ and $w \in L$, then $L_w \subseteq L$. This readily implies that $A^*\mathcal{J}^+$ is not closed under complement, since, for example, the complement of $(a+b)^*a(a+b)^*$ does not have this property. Thus $\mathcal{J}^+$ is not a variety of languages. On the other hand, it does satisfy many of the properties of varieties of languages: It is closed under finite unions and intersections, quotients, and inverse images of morphisms between free monoids.

It turns out that each of these three examples admits an algebraic characterization in terms of classes that are very much like pseudovarieties. For our first example, in which membership of a word in a language is determined by the leftmost letter, the correct generalization of pseudovarieties was already known to Eilenberg: One looks not at the syntactic monoid of a language $L$, but at the image of the set $A^+$ of nonempty words under the syntactic morphism. This is called the *syntactic semigroup* of $L$. We can define *pseudovarieties of finite semigroups* just as we defined pseudovarieties of finite monoids. Then $L \in A^*\mathcal{K}_1$ if and only if its syntactic semigroup belongs to the pseudovariety of semigroups defined by the identity $xy = x$. While $\mathcal{K}_1$ is not closed under inverse images of morphisms between free monoids, it is closed if we restrict ourselves to *non-erasing* morphisms—those that map every letter to a nonempty word.

We can use a similar method to characterize the class $\mathcal{QA}$. Once again we look not just at the syntactic monoid of a language $L$, but at the additional structure provided by the syntactic morphism $\eta_L$. It is known that $L \in A^*\mathcal{QA}$ if and only if for every $k \geqslant 0$, $\eta_L(A^k)$ contains no nontrivial groups [10]. The family $\mathbf{QA}$ of morphisms from free monoids onto finite monoids with this property forms a kind of pseudovariety with respect to appropriately modified definitions of direct product and division. An equational

characterization of **QA** is provided by the identity

$$(x^{\omega-1}y)^{\omega} = (x^{\omega-1}y)^{\omega+1},$$

where the identity is interpreted in the following sense: $\varphi \in \mathbf{QA}$ if and only if for all words $u$ and $v$ *of the same length*, $x = \varphi(u)$ and $y = \varphi(v)$ satisfy the identity. $\mathcal{QA}$ is closed under inverse images of morphisms $f\colon B^* \to A^*$ such that $f(B) \subseteq A^k$ for some $k \geqslant 0$; these are called *length multiplying morphisms*. In fact, these last two examples are instances of a single phenomenon: Families of morphisms $\varphi\colon A^* \to M$ onto finite monoids that form pseudovarieties with respect to some underlying composition-closed class $\mathcal{C}$ of morphisms between free monoids.

For the example $\mathcal{J}^+$ of $\Sigma_1$-definable languages, the algebraic characterization involves a different generalization of pseudovarieties. Here the additional structure on the syntactic monoid is provided by the embedding of $\eta_L(L)$ in $\mathrm{Synt}(L)$ : If $m_1, m_2 \in M$ then we say $m_1 \leqslant_L m_2$ if

$$\{(s,t) \in \mathrm{Synt}(L)\times\mathrm{Synt}(L) \mid sm_2t \in \eta_L(L)\} \subseteq \{(s,t) \in \mathrm{Synt}(L)\times\mathrm{Synt}(L) \mid sm_1t \in \eta_L(L)\}.$$

This gives a partial order on $\mathrm{Synt}(L)$ compatible with multiplication (see Section 4.4 in Chapter 1). We then find that $L \in A^*\mathcal{J}^+$ if and only if this ordered syntactic monoid satisfies the *inequality* $x \leqslant 1$ for each element $x$. The family of partially-ordered monoids satisfying this inequality is a pseudovariety of ordered finite monoids—it is closed under finite direct products, and order-compatible submonoids and quotients. The theory of pseudovarieties of ordered monoids and the corresponding *positive varieties* of languages is due to Pin [33]

In the next section we will formally develop the framework that gives the correspondence between pseudovarieties and language varieties, and the definition by profinite identities, in a very general setting. Pseudovarieties of finite monoids, as well as all the generalizations mentioned above, will appear as special cases.

# 2 Equations, identities and families of languages

The original statement of Eilenberg's theorem dealt exclusively with varieties of languages. Here we will show how to use a whole hierarchy of increasingly complex equational characterizations of increasingly structured families of languages. Before we describe these results, we need to give a quick introduction to the free profinite monoid and its connection to the theory of regular languages

## 2.1 The free profinite monoid

Say that a finite monoid $M$ *separates* two words $u, v \in A^*$ if there exists a morphism $\varphi\colon A^* \to M$ such that $\varphi(u) \neq \varphi(v)$. Note that if $u \neq v$, there always exists such a monoid. Indeed, for each $n \geqslant 1$, consider the quotient monoid $A^*/A^{\geqslant n}$: it consists of the set of words of length less than $n$, plus a zero, and each product with length at least $n$ (in $A^*$) is equal to 0. Then $A^*/A^{\geqslant n}$ separates $u$ and $v$ if $n > \max(|u|, |v|)$. We denote

by $r(u,v)$ the minimum cardinality of a monoid separating $u$ and $v$.

The *profinite distance* on $A^*$ is defined by letting $d(u,v) = 2^{-r(u,v)}$ if $u \neq v$ and $d(u,u) = 0$. One verifies easily that $d$ is in fact an ultrametric distance (it satisfies the ultrametric inequality $d(u,v) \leqslant \max(d(u,w), d(v,w))$, stronger than the triangle inequality), and the above discussion shows that the resulting metric space is Hausdorff.

The topology thus defined on $A^*$ is not especially interesting: we get a discrete space, where a sequence $(u_n)_n$ converges to a word $u$ if and only if $(u_n)_n$ is ultimately equal to $u$... This can be verified using the monoids $A^*/A^{\geqslant n}$ described above. There are, however, non-trivial Cauchy sequences. In fact, one can show the following.

**Proposition 2.1.** *A sequence $(u_n)_n$ is Cauchy if and only if, for each morphism $\varphi\colon A^* \to M$ into a finite monoid, the sequence $(\varphi(u_n))_n$ is ultimately constant.*

For instance, if $u$ is a word, then $(u^{n!})_n$ is a Cauchy sequence (this can be deduced from the fact that its image under any morphism into a finite monoid is ultimately constant), but it is non-trivial if $u \neq 1$. In topological terms, the uniform structure defined by the profinite distance is non-trivial.

Using a classical construction from topology (analogous to the construction of the real numbers from the rationals), we can now consider the completion of $(A^*, d)$, denoted by $\widehat{A^*}$. It can be viewed as the quotient of the set of Cauchy sequences in $(A^*, d)$ by the relation identifying two sequences $(u_n)$ and $(v_n)$ if the mixed sequence, alternating the terms of $(u_n)$ and $(v_n)$, is Cauchy as well. In particular, $A^*$ is naturally seen as a dense subset of $\widehat{A^*}$.

The following results can be verified by elementary means.

**Proposition 2.2.** *Let $A$ be an alphabet.*

(1) *The multiplication operation $(u,v) \mapsto uv$ in $A^*$ is uniformly continuous.*
(2) *Every morphism $\varphi\colon A^* \to B^*$ between free monoids, and every morphism $\psi\colon A^* \to M$ from a free monoid to a finite monoid (equipped with the discrete distance) is uniformly continuous.*
(3) *$\widehat{A^*}$ is a compact space.*

By a standard property of completions, it follows from Proposition 2.2 (1) that the multiplication of $A^*$ can be extended to $\widehat{A^*}$: the resulting monoid is called the *free profinite monoid on $A$*. Similarly, Proposition 2.2 (3) shows that each morphism $\varphi\colon A^* \to B^*$ between free monoids (resp. each morphism $\psi\colon A^* \to M$ from a free monoid to a finite monoid) admits a uniquely defined continuous extension, $\hat\varphi\colon \widehat{A^*} \to \widehat{B^*}$ (resp. $\hat\psi\colon \widehat{A^*} \to M$).

For example, consider the Cauchy sequence $(u^{n!})_n$, where $u \in A^*$, which we discussed above. This represents an element of $\widehat{A^*}$, which we will denote $u^\omega$. Observe that for any morphism $\varphi$ from $A^*$ into a finite monoid, the sequence $\hat\varphi(u^{n!})$ is ultimately constant and equal to the unique idempotent power of $\varphi(u)$, so in the notation we introduced earlier we have, very conveniently,

$$\hat\varphi(u^\omega) = (\varphi(u))^\omega.$$

We can similarly define $u^{\omega-1}$ as the element of $\widehat{A^*}$ represented by the Cauchy sequence $\hat{\varphi}(u^{n!-1})$.

Finally, we note the strong connection between regular languages and free profinite monoids.

**Proposition 2.3.** *Let $A$ be an alphabet and let $L \subseteq A^*$.*

(1) *$L$ is regular if and only if its topological closure in $\widehat{A^*}$, $\overline{L}$, is clopen (i.e., open and closed), if and only if $L = K \cap A^*$ for some clopen set $K \subseteq \widehat{A^*}$.*

(2) *If $L$ is regular and $u \in \widehat{A^*}$, then the following are equivalent:*
   - *(i)* $u \in \overline{L}$*;*
   - *(ii)* $\hat{\varphi}(u) \in \varphi(L)$ *for every morphism $\varphi$ from $A^*$ to a finite monoid;*
   - *(iii)* $\hat{\varphi}(u) \in \varphi(L)$ *for every morphism $\varphi$ from $A^*$ to a finite monoid recognizing $L$;*
   - *(iv)* $\hat{\eta}(u) \in \eta(L)$ *where $\eta$ is the syntactic morphism of $L$.*

## 2.2 Equations and lattices of languages

We begin our study of families of regular languages with the simplest such family: a lattice of languages over a fixed alphabet. In this chapter, we define a *lattice of languages* over an alphabet $A$ to be a set of languages over $A$ which is closed under finite union and finite intersection, and which contains $A^*$ and $\emptyset$ (respectively, the union and the intersection of an empty family of languages).

A *profinite equation on* $A$ is a pair $(u, v)$ of elements of $\widehat{A^*}$, usually denoted by $u \to v$. If $u, v \in A^*$, the equation is called *explicit*. A language $L \subseteq A^*$ is said to *satisfy* the equation $u \to v$, written $L \vdash u \to v$, if

$$u \in \overline{L} \Longrightarrow v \in \overline{L}.$$

**Remark 2.4.** It is important to note that $u$, $v$ and the words in $L$ are all defined over the same alphabet $A$. In contrast to the identities we encountered in Section 1, in this definition, the letters occurring in $u$ and $v$ are *not* considered as variables, to be replaced by arbitrary elements. We will formally define identities in Section 2.4.

The notion of equation is particularly relevant for regular languages. The following results directly from Proposition 2.3.

**Proposition 2.5.** *Let $L \subseteq A^*$ be regular and let $u, v \in \widehat{A^*}$.*

(1) *If $u, v \in A^*$, then $L \vdash u \to v$ if and only if $u \in L \Longrightarrow v \in L$.*

(2) *If $\eta$ is the syntactic morphism of $L$, then $L \vdash u \to v$ if and only if $\hat{\eta}(u) \in \eta(L) \Longrightarrow \hat{\eta}(v) \in \eta(L)$.*

Let $E$ be a set of equations on $A$. We denote by $\mathcal{L}(E)$ the set of regular languages in $A^*$ which satisfy all the equations in $E$. It is immediately verified that this set is closed under unions and intersections. Further, both $\emptyset$ and $A^*$ satisfy every equation. So $\mathcal{L}(E)$ is a lattice. The main theorem of this section states that all lattices of regular languages arise this way.

**Theorem 2.6.** *Let $\mathcal{L}$ be a class of regular languages in $A^*$. Then $\mathcal{L}$ is a lattice if and only if there exists a set $E$ of profinite equations on $A$ such that $\mathcal{L} = \mathcal{L}(E)$.*

We have already seen that one direction of this equivalence holds: every set of the form $\mathcal{L}(E)$ is a lattice. The proof of the converse is obtained after several steps. The first concerns the set of equations satisfied by a given language. If $L \subseteq A^*$, let

$$E_L = \left\{ (u, v) \in \widehat{A^*} \times \widehat{A^*} \mid L \vdash u \to v \right\}.$$

**Lemma 2.7.** *If $L$ is regular, then $E_L$ is clopen.*

*Proof.* By definition of the satisfaction of equations, we have

$$E_L = \left\{ (u, v) \in \widehat{A^*} \times \widehat{A^*} \mid (u \notin \overline{L}) \vee (v \in \overline{L}) \right\} = \left( \overline{L}^c \times \widehat{A^*} \right) \cup \left( \widehat{A^*} \times \overline{L} \right).$$

Lemma 2.7 follows from the fact that $\widehat{A^*}$, $\overline{L}$ and $\overline{L}^c$ are compact (since $L$ is regular). $\square$

The proof of the next claim illustrates the crucial role played by the compactness of $\widehat{A^*}$. Let $\mathcal{L}$ be a lattice of regular languages in $A^*$ and let $E_{\mathcal{L}} = \bigcap_{L \in \mathcal{L}} E_L$.

**Lemma 2.8.** *Let $L$ be a regular language in $\mathcal{L}(E_{\mathcal{L}})$: that is, $L$ satisfies all the profinite equations satisfied by all the elements of $\mathcal{L}$. Then there exists a finite subset $\mathcal{K}$ of $\mathcal{L}$ such that $L \in \mathcal{L}(E_{\mathcal{K}})$.*

*Proof.* By Lemma 2.7, $E_L$ and each $E_K^c$ ($K \in \mathcal{L}$) are open sets. Moreover, if $(u, v)$ does not belong to any of the $E_K^c$ ($K \in \mathcal{L}$), then $(u, v)$ belongs to each $E_K$, that is, every language in $\mathcal{L}$ satisfies $u \to v$. It follows that $L$ satisfies $u \to v$ as well, that is, $(u, v) \in E_L$. Therefore $E_L$ and the $E_K^c$ ($K \in \mathcal{L}$) form an open cover of $\widehat{A^*}$.

By compactness, there exists a finite subcollection $\mathcal{K}$ of $\mathcal{L}$ such that $\widehat{A^*}$ is covered by $E_L$ and the $E_K^c$, $K \in \mathcal{K}$. It follows that $E_L$ contains the complement of $\bigcup_{K \in \mathcal{K}} E_K^c$, namely the intersection $\bigcap_{K \in \mathcal{K}} E_K$. That is, $L$ satisfies all the equations satisfied by the elements of $\mathcal{K}$, which establishes the claim. $\square$

We are now ready to prove Theorem 2.6, by showing that if $\mathcal{L}$ is a lattice of regular languages in $A^*$, then $\mathcal{L} = \mathcal{L}(E_{\mathcal{L}})$. It is immediate by construction that $\mathcal{L}$ is contained in $\mathcal{L}(E_{\mathcal{L}})$. Let us now consider a language $L \in \mathcal{L}(E_{\mathcal{L}})$. By Lemma 2.8, we have $L \in \mathcal{L}(E_{\mathcal{K}})$ for a finite subset $\mathcal{K}$ of $\mathcal{L}$.

For each $u \in L$, let $\mathcal{K}(u)$ be the intersection of the languages $K \in \mathcal{K}$ containing $u$. Even though $L$ may be infinite, $\mathcal{K}(u)$ takes only finitely many values since $\mathcal{K}$ is finite. By definition of the $\mathcal{K}(u)$, we have $L \subseteq \bigcup_{u \in L} \mathcal{K}(u)$, a finite union.

Conversely, let $v \in \bigcup_{u \in L} \mathcal{K}(u)$. Then there exists a word $u \in L$ such that $v$ belongs to every $K \in \mathcal{K}$ containing $u$. That is, every $K \in \mathcal{K}$ satisfies the equation $u \to v$. In other words, $u \to v$ lies in $E_{\mathcal{K}}$, and hence $L$ satisfies that equation. Since $u \in L$, it follows that $v \in L$. Thus $L = \bigcup_{u \in L} \mathcal{K}(u)$ and hence $L \in \mathcal{L}$, which concludes the proof.

## 2.3 More classes of languages: from lattices to varieties

Here we explore how classes of regular languages that are more structured than lattices can be defined by more structured sets of equations. We start with an elementary lemma.

**Lemma 2.9.** *Let $\mathcal{L}$ be a lattice of regular languages satisfying the profinite equation $u \to v$.*

(1) *If $\mathcal{L}$ is closed under complementation, then $\mathcal{L}$ also satisfies $v \to u$.*
(2) *If $\mathcal{L}$ is closed under quotients, then $\mathcal{L}$ satisfies the equations $xuy \to xvy$, for all $x, y \in \widehat{A^*}$.*

*Proof.* It follows from the definition of equations that $L$ satisfies $u \to v$ if and only if its complement satisfies $v \to u$. The first part of the claim follows immediately.

It is also elementary that, if $x, y \in A^*$ and $x^{-1}Ly^{-1} \vdash u \to v$, then $L \vdash xuy \to xvy$. Thus, if $\mathcal{L}$ is closed under quotients, then $\mathcal{L}$ satisfies all the equations $xuy \to xvy$ with $x, y \in A^*$. This holds also if $x, y \in \widehat{A^*}$ since $E_{\mathcal{L}}$ is closed and $A^*$ is dense in $\widehat{A^*}$. $\square$

We now extend the notion of profinite equations as follows: if $u, v \in \widehat{A^*}$, we say that a language $L$ satisfies the *symmetrical* equation $u \leftrightarrow v$ if $L$ satisfies both $u \to v$ and $v \to u$.

We also say that a language $L$ satisfies the *profinite inequality* $v \leqslant u$ if it satisfies all the equations of the form $xuy \to xvy$ with $x, y \in \widehat{A^*}$, and it satisfies the *profinite equality* $u = v$ if if satisfies both $u \leqslant v$ and $v \leqslant u$. The verification of the following corollary is now elementary.

**Corollary 2.10.** *Let $\mathcal{L}$ be a set of regular languages in $A^*$.*

(1) *Then $\mathcal{L}$ is a boolean algebra if and only if $\mathcal{L} = \mathcal{L}(E)$ for some set $E$ of symmetrical profinite equations on $A$.*
(2) *$\mathcal{L}$ is a lattice closed under quotients if and only if $\mathcal{L} = \mathcal{L}(E)$ for some set $E$ of profinite inequalities on $A$.*
(3) *$\mathcal{L}$ is a boolean algebra closed under quotients if and only if $\mathcal{L} = \mathcal{L}(E)$ for some set $E$ of profinite equalities on $A$.*

## 2.4 Identities and varieties

We now come to the historically and mathematically important class of varieties. Varieties of languages were defined in Section 1.3 but we will not use this definition here. In fact, in the course of this section, we will give an alternate, equivalent definition of varieties.

An important difference between varieties and the lattices of languages over a fixed alphabet discussed so far in Section 2, is that a variety $\mathcal{V}$ consists of a collection of lattices $A^*\mathcal{V}$, one for each finite alphabet $A$. More generally, we define a *class of regular languages* $\mathcal{V}$ to be an operator which assigns to each finite alphabet $A$, a family $A^*\mathcal{V}$ of regular languages in $A^*$.

First, we prove a technical lemma.

**Lemma 2.11.** *Let $\varphi\colon A^* \to B^*$ be a morphism, $L \subseteq B^*$ and $u, v \in \widehat{A^*}$.*

(1) $\hat\varphi(u) \in \overline{L}$ *if and only if* $u \in \overline{\varphi^{-1}(L)}$.
(2) $L$ *satisfies* $\hat\varphi(u) \to \hat\varphi(v)$ *if and only if* $\varphi^{-1}(L)$ *satisfies* $u \to v$.

*Proof.* The first statement is trivial if $u, v \in A^*$: indeed, $\varphi$ and $\hat\varphi$ coincide on words, and the intersection of $\overline{L}$ (resp. $\overline{\varphi^{-1}(L)}$) with $A^*$ (resp. $B^*$) is $L$ (resp. $\varphi^{-1}(L)$). The extension to the case where $u, v \in \widehat{A^*}$ is obtained by density.

The second statement follows immediately from the first and the definition of profinite equations. □

We extend the notion of profinite equations, this time to profinite identities, to permit the treatment of classes of regular languages instead of lattices of regular languages over a fixed alphabet. Since there is no alphabet of reference anymore, we will usually denote by $X$ the alphabet over which profinite identities are written.

Let $\mathcal{C}$ be a composition-closed class of morphisms between free monoids, $u, v \in \widehat{X^*}$ and $L \subseteq A^*$, where $X$ and $A$ are finite, but possibly different alphabets. We say that $L$ *$\mathcal{C}$-identically satisfies* $u \to v$ if, for each morphism $\varphi\colon X^* \to A^*$ in $\mathcal{C}$, $L$ satisfies $\hat\varphi(u) \to \hat\varphi(v)$. We say that a class of regular languages $\mathcal{V}$ $\mathcal{C}$-identically satisfies an equation if $A^*\mathcal{V}$ does, for each finite alphabet $A$.

The following statement is a direct application of Lemma 2.11.

**Corollary 2.12.** *Let $\mathcal{V}$ be a class of regular languages, let $\mathcal{C}$ be a family of morphisms between free monoids closed under composition, such that whenever $\varphi\colon A^* \to B^*$ is in $\mathcal{C}$ and $L \in B^*\mathcal{V}$, then $\varphi^{-1}(L) \in A^*\mathcal{V}$.*

*If $X^*\mathcal{V}$ satisfies the profinite equation $u \to v$ (with $u, v \in \widehat{X^*}$), then $\mathcal{V}$ $\mathcal{C}$-identically satisfies $u \to v$.*

Using the notions introduced in Section 2.3, we say that $L$ satisfies the *profinite $\mathcal{C}$-identity* $u = v$ (resp. *profinite ordered $\mathcal{C}$-identity* $u \leqslant v$) if $L$ $\mathcal{C}$-identically satisfies $u = v$ (resp. $u \leqslant v$). If $E$ is a set of profinite equations and for each finite alphabet $A$, $A^*\mathcal{V}$ is the set of regular languages in $A^*$ which $\mathcal{C}$-identically satisfy the elements of $E$, we say that the resulting class of regular languages $\mathcal{V}$ is *$\mathcal{C}$-defined by $E$*.

Let us now define (positive) $\mathcal{C}$-varieties: a class $\mathcal{V}$ of regular languages is a *positive $\mathcal{C}$-variety* (resp. a *$\mathcal{C}$-variety*) *of languages* if each $A^*\mathcal{V}$ is a lattice (resp. a boolean algebra) closed under quotients and if, for each $\varphi\colon A^* \to B^*$ in $\mathcal{C}$ and each $L \in B^*\mathcal{V}$, we have $\varphi^{-1}(L) \in A^*\mathcal{V}$.

If $\mathcal{C}$ is the class of all morphisms between free monoids, we drop the prefix $\mathcal{C}$ and simply talk of (ordered) profinite identities and (positive) varieties of languages.

Collecting Corollaries 2.10 and 2.12, we have the following characterizations.

**Theorem 2.13.** *Let $\mathcal{V}$ be a class of regular languages and let $\mathcal{C}$ be a composition-closed class of morphisms between free monoids. Then $\mathcal{V}$ is a positive $\mathcal{C}$-variety (resp. a $\mathcal{C}$-variety) if and only if $\mathcal{V}$ is $\mathcal{C}$-defined by a set of profinite ordered $\mathcal{C}$-identities (resp. profinite $\mathcal{C}$-identities).*

**Remark 2.14.** In Section 1.3, we gave a different definition of varieties of languages, and Theorem 1.3 stated that it was equivalent to the definition given above. We will prove this equivalence in Section 2.5 below, thus formally reconciling the two definitions.

### 2.5 Eilenberg's and Reiterman's theorems

We note that (in)equalities can be interpreted in the (ordered) syntactic monoid of a language. Let $L$ be a regular language in $A^*$ and let $u, v \in \widehat{A^*}$. By Proposition 2.5, if $\eta$ is the syntactic morphism of $L$, then $L \vdash v \leqslant u$ if and only if $\hat{\eta}(v) \leqslant_L \hat{\eta}(u)$.

Thus membership of a regular language $L$ in a lattice of regular languages closed under quotients is characterized by properties of the syntactic morphism of $L$.

We can also interpret identities in abstract finite ordered monoids—that is, finite monoids in which there is a partial order $\leqslant$ compatible with multiplication: If $u, v \in \widehat{X^*}$, we say that a finite ordered monoid $M$ satisfies the profinite identity $u \leqslant v$ if for every morphism $\varphi\colon X^* \to M$ we have $\hat{\varphi}(u) = \hat{\varphi}(v)$. Likewise a monoid $M$ satisfies the profinite identity $u = v$ if for each such $\varphi$ we have $\hat{\varphi}(u) = \hat{\varphi}(v)$. We extend this notion further to $\mathcal{C}$-satisfaction of identities. We call a morphism $\varphi\colon A^* \to M$, where $M$ is finite and $\varphi$ maps onto $M$, a *stamp*. We also define *ordered stamps* as morphisms from a free monoid $A^*$ onto an ordered finite monoid. (Such morphisms are automatically order-preserving if we consider the trivial ordering on $A^*$ in which $w_1 \leqslant w_2$ if and only if $w_1 = w_2$.) Let $\mathcal{C}$ be a class of morphisms between finitely generated free monoids that is closed under composition and that contains all the length-preserving morphisms. We say that the ordered stamp $\varphi\colon A^* \to (M, \leqslant)$ *$\mathcal{C}$-satisfies the profinite identity* $u \leqslant v$ with $u, v \in \widehat{X^*}$ if and only if for all morphisms $\psi\colon X^* \to A^*$ with $\psi \in \mathcal{C}$, we have $\hat{\varphi}\hat{\psi}(u) \leqslant \hat{\varphi}\hat{\psi}(v)$. We similarly define $\mathcal{C}$-satisfaction of identities $u = v$ by (not necessarily ordered) stamps.

We have already defined pseudovarieties of finite monoids in Section 1. We can extend this definition to define $\mathcal{C}$-pseudovarieties of stamps. We call a collection **V** of stamps a *$\mathcal{C}$-pseudovariety* if it satisfies the following two conditions:

(i) If $\varphi\colon A^* \to M$ is in **V**, $\psi\colon B^* \to A^*$ is in $\mathcal{C}$, and $\eta$ is a morphism from $Im(\varphi\psi)$ onto a finite monoid $N$, then $\eta\varphi\psi\colon B^* \to N$ is in **V**.

(ii) If $\varphi_i\colon A^* \to M_i$ are in **V** for $i = 1, 2$, then $\varphi_1 \times \varphi_2\colon A^* \to Im(\varphi_1 \times \varphi_2) \subseteq M_1 \times M_2$ is in **V**.

If we restrict the morphisms occurring in these definitions to order-preserving morphisms or ordered monoids, we obtain the definition of *ordered $\mathcal{C}$-pseudovarieties of stamps*. Ordinary pseudovarieties coincide with $\mathcal{C}$-pseudovarieties in the case where $\mathcal{C}$ contains all morphisms between finitely-generated free monoids.

We say that a class **V** of finite (ordered) monoids is *defined* by a set $E$ of identities (written $\mathbf{V} = [\![E]\!]$) if **V** consists of all the finite (ordered)monoids that satisfy all of the identities in $E$. Similarly, we say that a family **V** of stamps is $\mathcal{C}$-defined by $E$ (we write $\mathbf{V} = [\![E]\!]_{\mathcal{C}}$) if **V** consists of all the stamps that $\mathcal{C}$-satisfy these identities.

Further if **V** is a class of monoids or stamps, ordered or unordered, we define the corresponding class $\mathcal{V}$ of languages by setting $L \in A^*\mathcal{V}$ if and only if $\mathrm{Synt}(L) \in \mathbf{V}$ (if **V** is a class of monoids) or $\eta_L \in \mathbf{V}$ (if **V** is a class of stamps). We write $\mathbf{V} \mapsto \mathcal{V}$ to denote this correspondence.

This leads us to a restatement of Eilenberg's Theorem, Theorem 1.3 above, as well as its generalization to $\mathcal{C}$-varieties, and allows us to prove it simultaneously with Reiterman's Theorem.

**Theorem 2.15.**

(1) (Eilenberg's Theorem) *If* **V** *is a pseudovariety (respectively* $\mathcal{C}$*-pseudovariety, ordered pseudovariety) and* **V** $\mapsto \mathcal{V}$*, then* $\mathcal{V}$ *is a variety of languages (respectively* $\mathcal{C}$*-variety of languages, positive variety of languages) and in each case this gives a one-to-one correspondence between pseudovarieties and varieties of languages.*

(2) (Reiterman's Theorem) *A class* **V** *of monoids (stamps, ordered monoids) is a pseudovariety (respectively* $\mathcal{C}$*-pseudovariety, ordered pseudovariety) if and only if it is defined (*$\mathcal{C}$*-defined) by a set of profinite identities.*

In the argument we sketch below, we confine ourselves to the case of ordinary monoids, but everything generalizes in an entirely straightforward fashion to ordered monoids and stamps. The key to the proofs of both parts of the theorem is Theorem 2.13 above, along with the following elementary but very useful lemma, already brought to the reader's attention in Section 1.1.

**Lemma 2.16.** *Let* $\varphi\colon A^* \to M$ *be a morphism into a finite monoid. Then* $M$ *divides the direct product of the syntactic monoids of the syntactic monoids of the languages* $\varphi^{-1}(m)$*,* $m \in M$*.*

*Proof.* For each $m \in M$, let $\eta_m\colon A^* \to \mathrm{Synt}(\varphi^{-1}(m))$ be the syntactic morphism of $\varphi^{-1}(m)$. It suffices to show that for each $u, v \in A^*$, $\eta_m(u) = \eta_m(v)$ for each $m \in M$ implies $\varphi(u) = \varphi(v)$.

Indeed, let $m = \varphi(u)$. Then $u \in \varphi^{-1}(m)$ and since $\eta_m(v) = \eta_m(u)$, we have $v \in \varphi^{-1}(m)$, $\varphi(v) = m = \varphi(u)$. □

**Corollary 2.17.** *Every pseudovariety of monoids is generated by the syntactic monoids it contains.*

*Proof.* The result follows directly from Lemma 2.16, since $M$ recognizes each $\varphi^{-1}(M)$ ($m \in M$): thus each $\mathrm{Synt}(\varphi^{-1}(m))$ divides $M$ and hence lies in the pseudovarieties containing $M$. □

Now let $\mathcal{V}$ be a variety of languages and let $E$ be a set of profinite identities defining $\mathcal{V}$. Let also **V** be the class of finite monoids satisfying the profinite identities in $E$. It is easily verified that **V** is a pseudovariety.

Moreover, if $L$ is a regular language in $A^*$, we have $L \in A^*\mathcal{V}$ if and only if $L \vdash E$, if and only if $\mathrm{Synt}(L)$ satisfies the profinite identities in $E$, if and only if $\mathrm{Synt}(L) \in$ **V**.

Thus **V** $\mapsto \mathcal{V}$ in the correspondence described in Section 1.3. If **W** is another pseudovariety such that **W** $\mapsto \mathcal{V}$, then **V** and **W** contain the same syntactic monoids, and Corollary 2.17 shows that **V** $=$ **W**. This establishes Eilenberg's Theorem.

For Reiterman's Theorem, we start with a pseudovariety **V** and consider the associated variety of languages $\mathcal{V}$. The above reasoning shows that **V** is defined by any set of profinite identities which, seen in the setting of classes of languages, defines $\mathcal{V}$.

Note that these proofs are different from the classical proofs of Eilenberg's theorem, in [18] or [32], and of Reiterman's theorem, in [3], [36] or [38].

## 2.6 Examples of varieties

We now look at some concrete instances of varieties, revisiting our examples from Section 1, among others, in light of the theory presented above. In doing so, we will work from both sides of the correspondence between pseudovarieties and varieties of languages, at times beginning with a variety of languages, at others with a property of a class of finite monoids.

**2.6.1 Idempotent and commutative monoids** We begin, as before, with the variety of languages corresponding to the pseudovariety $\mathbf{J}_1$. For each finite alphabet $A$, let $A^*\mathcal{J}_1$ be the smallest boolean-closed family of subsets of $A^*$ that contains all the languages $B^*$, where $B \subseteq A$. Equivalently, it is the smallest boolean-closed set containing all the $A^*aA^*$ ($a \in A$). Putting it again differently, $A^*\mathcal{J}_1$ is precisely the family of languages $L$ in $A^*$ for which membership of a word $w$ in $L$ depends only on the set $\alpha(w)$ of letters of $w$. This is because

$$\{v \in A^* \mid \alpha(v) = \alpha(w)\} = \alpha(w)^* \setminus \bigcup_{B \subsetneq \alpha(w)} B^*.$$

Observe that for all $a \in A$ and $B \subseteq A$,

$$a^{-1}B^* = B^*a^{-1} = \begin{cases} \emptyset & \text{if } a \notin B \\ B^* & \text{if } a \in B. \end{cases}$$

Further, if $C$ is another finite alphabet and $\varphi\colon C^* \to A^*$ is a morphism,

$$\varphi^{-1}(B^*) = (C \cap \varphi^{-1}(B))^*.$$

Left and right quotient and inverse image under morphisms all commute with boolean operations. So these two observations imply, independently of any algebraic considerations, that $\mathcal{J}_1$ is a variety of languages, and thus, by Theorem 2.13 is defined by a set of profinite identities. Further, from our proof of Eilenberg's Theorem, the same set of identities defines the corresponding pseudovariety of finite monoids.

Of course, we have already exhibited these identities, but let us see what they look like in the context of our equational theory. Let $X = \{x, y\}$, and let $A$ be any finite alphabet. Every language $L \in A^*\mathcal{J}_1$ satisfies the identities $xy = yx$ and $x^2 = x$, since for any morphism $\varphi\colon X^* \to A^*$ and any $u, v \in A^*$, $\alpha(u\varphi(xy)v) = \alpha(u\varphi(yx)v)$, and $\alpha(u\varphi(x^2)v) = \alpha(u\varphi(x)v)$. Conversely, suppose $L \subseteq A^*$ satisfies these identities. We will show $L \in A^*\mathcal{J}_1$: Let $w, w' \in B^*$, with $w \in L$ and $\alpha(w) = \alpha(w')$. We claim $w' \in L$. Since $\alpha(w) = \alpha(w')$, we can transform both $w$ and $w'$ into a common normal form $w''$ by successively interchanging adjacent letters until the word is sorted (with respect to some total ordering on $A$) and then replacing occurrences of $aa$ by $a$, where $a \in A$. Interchanging adjacent letters entails replacing $ua_1a_2v$ by $ua_2a_1v$, where $u, v \in A^*$ and $a_1, a_2 \in A$. Since $L$ satisfies the identity $xy = yx$, if $ua_1a_2v \in L$ then $ua_2a_1v \in L$ (using the morphism $\varphi\colon X^* \to A^*$ that maps $x, y$ to $a_1, a_2$, respectively.). Similarly, replacing $aa$ by $a$ preserves membership in $L$, since $L$ satisfies the identity $x^2 = x$. Thus $\mathcal{J}_1$ is defined by this pair of identities. It follows that the corresponding pseudovariety $\mathbf{J}_1$ of finite monoids is defined by the same pair of identities, and thus consists of the idempotent and commutative monoids.

**2.6.2 Piecewise-testable languages** Now let us consider the piecewise-testable languages of Section 1.2. We denote the family of piecewise-testable languages over a finite alphabet $A$ by $A^*\mathcal{J}$. Let us look at the profinite identities satisfied by these languages. As observed earlier (Section 2.1), if $u \in X^*$ then the sequence $(u^{n!})_n$ is a Cauchy sequence whose limit is written $u^\omega$. Moreover, for any morphism $\varphi\colon X^* \to A^*$, where $A$ is a finite alphabet, $\hat{\varphi}(u^\omega) = (\hat{\varphi}(u))^\omega$ (the idempotent power of $\hat{\varphi}(u)$). Now let $X = \{x, y\}$. We claim that every piecewise-testable language $L$ over $A^*$ satisfies the profinite identities

$$(xy)^\omega x = (xy)^\omega = y(xy)^\omega.$$

This is equivalent to saying that for all $s, t, u, v \in A^*$,

$$s(tu)^\omega tv \in \overline{L} \Leftrightarrow s(tu)^\omega v \in \overline{L} \Leftrightarrow su(tu)^\omega v \in \overline{L}.$$

Now fix an integer $k > 0$. For sufficiently large values of $n$, the words

$$s(tu)^{n!}tv, s(tu)^{n!}v, su(tu)^{n!}v$$

contain the same subwords of length $k$. Since $L$ is piecewise-testable, for sufficiently large $n$, all but finitely many of the terms of the three sequences are either all in $L$ or all outside of $L$. Since $\overline{L}$ is clopen, the three respective limits are either all in $\overline{L}$ or all outside $\overline{L}$.

Thus, as we showed in Section 2.5, the syntactic monoid of any piecewise testable language satisfies these same profinite identities. We arrive again at the observation that the syntactic monoid of every piecewise-testable language satisfies the identities $(xy)^\omega x = (xy)^\omega = y(xy)^\omega$. That these identities define the pseudovariety **J** of finite $\mathcal{J}$-trivial monoids is simple to establish. That they completely characterize the variety of piecewise-testable languages is the deep content of Simon's Theorem [43].

**2.6.3 Group languages** Similarly, the pseudovariety **G** of finite groups is defined by the profinite identity $x^\omega = 1$. As a consequence, the corresponding variety $\mathcal{G}$ of languages is defined by the same profinite identity. In contrast to the other examples presented here, we do not possess a simple description of $\mathcal{G}$ in terms of basic operations on words.

**2.6.4 Left-zero semigroups** We already appealed to Eilenberg's Theorem in Section 1 to show that the class $\mathcal{K}_1$ is not a variety of languages. But we can show here that it is a $\mathcal{C}$-variety for a slightly restricted class $\mathcal{C}$ of morphisms. Let $\mathcal{C}_{ne}$ denote the class of *non-erasing* morphisms between finitely-generated free monoids–those $\varphi\colon A^* \to B^*$ such that for all $a \in A$, $\varphi(a) \neq 1$. Let $L \in A^*\mathcal{K}_1$. If $s, t, u, v \in A^*$, and $t, u \neq 1$, then $stuv \in L$ if and only if $stv \in L$. Moreover, this property of $L$ characterizes membership in $A^*\mathcal{K}_1$. One way to state this property is that the variety of languages $\mathcal{K}_1$ is defined by the $\mathcal{C}_{ne}$-identity $xy = x$. Equivalently, the corresponding $\mathcal{C}_{ne}$-pseudovariety $\mathbf{K}_1$ of stamps is defined by the same $\mathcal{C}_{ne}$-identity. This means $(\varphi\colon A^* \to M) \in \mathbf{K}_1$ if $\varphi(uv) = \varphi(u)$, for $u, v \in A^+$.

Alternatively, one may consider, instead of the $\mathcal{C}_{ne}$-pseudovariety generated by the syntactic morphisms of languages in $\mathcal{K}_1$, the pseudovariety of finite *semigroups* generated by the images of nonempty words under the syntactic morphisms. This was the approach originally taken, but here we prefer to emphasize that all these many different flavors of pseudovarieties can be treated in the same general setting.

**2.6.5 Quasiaperiodic stamps** Whenever we have a morphism $\varphi\colon A^* \to M$, the family of sets

$$\{\varphi(A^s) \mid s > 0\}$$

forms a subsemigroup of the power set semigroup $\mathcal{P}(M)$. As this is a finite cyclic semigroup, generated by $\varphi(A)$, it contains a unique idempotent. Thus there is some $s > 0$ such that $\varphi(A^s) = \varphi(A^{2s})$, so that $\varphi(A^s)$ is a subsemigroup of $M$. We call this the *stable semigroup* of $\varphi$. Let $\mathbf{QA}$ denote the set of morphisms $\varphi$ from a free finitely-generated monoid onto a finite monoid such that $\varphi$ is surjective, and the stable semigroup of $\varphi$ is aperiodic.

We claim $\mathbf{QA}$ is a $\mathcal{C}_{lm}$-pseudovariety of stamps, where $\mathcal{C}_{lm}$ consists of morphisms $\psi\colon A^* \to B^*$ between finitely generated free monoids such that all $\psi(a)$, where $a \in A$, are nonempty words having the same length. (The letters *lm* stand for *length-multiplying*, since the lengths of all words in $A^*$ are multiplied by a constant factor when $\psi$ is applied.) To see this, suppose $(\varphi\colon B^* \to M) \in \mathbf{QA}$, and $(\psi\colon A^* \to B^*) \in \mathcal{C}_{lm}$. Let $\varphi(B^s)$ be the stable semigroup of $\varphi$, $\varphi\psi(A^t)$ the stable semigroup of $\varphi\psi\colon A^* \to Im(\varphi\psi)$, and $k$ the length of each $\psi(a)$ for $a \in A$. Then $\varphi\psi(A^t) = \varphi\psi(A^{st}) \subseteq \varphi(A^{kst}) = \varphi(A^s)$, and thus the stable semigroup of $\varphi\psi$ is also aperiodic. Further, if the stable semigroups $\varphi_j(A^{s_j})$ of stamps $\varphi_j\colon A^* \to M_j$, for $j = 1, 2$, are aperiodic, then the stable semigroup of $\varphi_1 \times \varphi_2$ is contained in $\varphi_1(A^{s_1}) \times \varphi_2(A^{s_2})$, and is therefore aperiodic.Thus $\mathbf{QA}$ is a $\mathcal{C}_{lm}$-pseudovariety, and is accordingly defined by a set of profinite $\mathcal{C}_{lm}$-identities. What does it mean for a stamp $\varphi\colon A^* \to M$ to satisfy a $\mathcal{C}_{lm}$ identity $u = v$? In such an identity, $u$ and $v$ are elements of $\widehat{X^*}$ for some finite alphabet $X$. The identity is satisfied if for every morphism $\psi\colon X^* \to A^*$ in $\mathcal{C}_{lm}$, $\hat{\varphi}\hat{\psi}(u) = \hat{\varphi}\hat{\psi}(v)$. Informally, this says that so long as we replace the letters in $u$ and $v$ by elements of $A^+$ that all have the same length, the images in $M$ are identical. We claim that **QA** is defined by the single profinite $\mathcal{C}_{lm}$-identity

$$(x^{\omega-1}y)^{\omega} = (x^{\omega-1}y)^{\omega+1}.$$

Let us prove this. First, we show that $\mathbf{QA}$ satisfies the identity. Let $(\varphi\colon A^* \to M) \in \mathbf{QA}$, and choose $p > 0$ such that for all $m \in M$, $m^p$ is idempotent. We then also have $m^{ps}$ idempotent for all $m \in M$, where $\varphi(A^s)$ is the stable semigroup of $\varphi$. If the identity is *not* satisfied, then there exist words $u$ and $v$ in $B^*$, both of length $k > 0$, such that

$$(\varphi(u^{ps-1}v))^{ps} \neq (\varphi(u^{ps-1}v))^{ps+1}.$$

Thus $\{(\varphi(u^{ps-1}y))^{ps+r} \mid r \geqslant 0\}$ is a nontrivial group in $\varphi((A^s)^+) = \varphi(A^s)$, contradicting membership in $\mathbf{QA}$. Conversely, suppose a stamp $\varphi\colon A^* \to M$ satisfies the identity. Suppose the stable semigroup $\varphi(A^s)$ contains a group element $g = \varphi(u)$, with $|u| = s$. Let $e = \varphi(v)$, where $|v| = s$ is the identity of this group. Since $\varphi$ satisfies the identity,

$$e = \varphi((u^{\omega-1}v)^{\omega}) = \varphi((u^{\omega-1}v)^{\omega+1}) = g^{-1},$$

so every group in $\varphi(A^s)$ is trivial.

We introduced the $\mathcal{C}_{lm}$-pseudovariety $\mathbf{QA}$ in Section 1 in quite different terms, by giving a logical description of the corresponding $\mathcal{C}_{lm}$-variety of languages. We will show in Section 3 that they do in fact correspond.

**2.6.6 $\Sigma_1$-languages** As in Section 1.2.2, we denote by $A^*\mathcal{J}^+$ the family of langues over $A$ defined by $\Sigma_1$ sentences. Languages in this family are precisely the finite unions of the languages $L_v$, where $v \in A^*$. We claim that $\mathcal{J}^+$ is defined by the profinite ordered identity $x \leqslant 1$. A language $L$ satisfies this identity if and only if for all $u, v, w \in A^*$, whenever $uw \in L$, then $uvw \in L$. Clearly, each $L_v$ satisfies this identity. We must show, conversely, that any language satisfying this identity is a finite union of $L_v$ for various $v \in A^*$. Certainly, if $L$ satisfies the identity and $v \in L$, then $L_v \subseteq L$, so that

$$L = \bigcup_{v \in L} L_v.$$

We need to show that this can be replaced by a finite union. Let $T$ consist of the *subword-minimal* elements of $L$, that is, those $v \in L$ such that no proper subword of $v$ is in $L$. Then

$$L = \bigcup_{v \in T} L_v.$$

We now invoke a theorem of G.Higman [24]: The subword ordering in $A^*$ has no infinite antichains: That is, any set $T$ of words in which no element is a strict subword of another element is finite.

The corresponding ordered pseudovariety $\mathbf{J}^+$ consequently consists of all partially ordered finite monoids for which the identity 1 is the maximum element, and thus a language belongs to $A^*\mathcal{J}^+$ if and only if its ordered syntactic monoid satisfies this property.

**2.6.7 Languages with zero** All of our examples so far have concerned some flavor of varieties of languages, language families that are defined across all finite alphabets and are closed under inverse images of morphisms between free monoids. Part of the great novelty of the equational theory of Gehrke *et al.* [22] presented here is that it applies to language classes with weaker closure properties. Here we give a simple example.

We say a regular language $L \subseteq A^*$ is a *language with zero* if $\mathrm{Synt}(L)$ has a zero. This is equivalent to saying that there is a two-sided ideal $J$ in $A^*$ such that either $J \subseteq L$ or $L \cap J = \emptyset$. This property is easily seen to be closed under boolean operations and quotients. It is, not, however, closed under inverse images of any composition-closed class $\mathcal{C}$ of morphisms that contains the length-preserving morphisms. Indeed, let $L \subseteq A^*$ be any regular language without a zero, and let $b$ be a new letter. Then, viewed as a subset of $(A \cup \{b\})^*$, $L$ has a zero, so this class is not closed under the inverse image of the length-preserving morphism that embeds $A^*$ in $(A \cup \{b\})^*$. Nonetheless, by our Corollary 2.10, this class of languages is defined by a set of profinite inequalities.

We now exhibit such a set of inequalities. We start by defining three sequences of words in $A^*$. Let

$$u_1, u_2, \ldots$$

be any enumeration of the elements of $A^*$, let

$$v_n = u_1 \cdots u_n,$$

and

$$w_1 = 1, w_{n+1} = (w_n v_n w_n)^{n!}.$$

Look at the image of the $w_i$ under a surjective morphism $\varphi\colon A^* \to M$, where $M$ is finite. Since every $u \in A^*$ occurs as a factor of all but finitely many $w_i$, almost all $\varphi(w_i)$ are in the minimal ideal $K$ of $M$. Since for all $m \in M$, $m^{n!}$ is idempotent for sufficiently large $n$, almost all $\varphi(w_i)$ are idempotents in the minimal ideal of $M$. Finally, if $\varphi(w_i)$ is such an idempotent $e$, then $\varphi(w_{i+1})$ is an idempotent in $eKe$, and so is itself equal to $e$. Thus for every finite monoid, the sequence $(\varphi(w_n))_n$ is convergent, so $(w_n)_n$ converges to an element $\rho_A$ of $\widehat{A^*}$, such that $\hat{\varphi}(\rho_A)$ is an idempotent in the minimal ideal of $\varphi(A^*)$.

Suppose $L \subseteq A^*$ has a zero. Then the minimal ideal of $\mathrm{Synt}(L)$ consists of this 0 alone, so if $\eta$ is the syntactic morphism of $L$ and $a \in A$, $\hat{\eta}(\rho_A) = \hat{\eta}(a\rho_A) = \hat{\eta}(\rho_A a)$. Thus $L$ satisfies the equalities

$$a\rho_A = \rho_A = \rho_A a$$

for all $a \in A$. Conversely, if $L$ satisfies these equalities, then the minimal ideal of $\eta(A^*)$ contains just one element, so $L$ is a language with zero. So these equalities define the class of languages with zero.

**2.6.8 Languages defined by density** Say that a language $L \subseteq A^*$ is *dense* if every word of $A^*$ occurs as a factor of a word in $L$, that is, $L \cap A^*uA^* \neq \emptyset$ for every $u \in A^*$. The set consisting of $A^*$ and the non-dense languages forms a quotient-closed lattice, which is defined by the profinite inequalities $x \leqslant 0$ ($x \in A^*$)—this is short for $a\rho_A = \rho_A a = \rho_A$ for every $a \in A$ and $x \leqslant \rho_A$ for every $x \in A^*$ [22].

Now define the *density of a language* $L$ as the function $d_L(n)$ which counts the number of words of length $n$ in $L$. A language with bounded density (also called *slender*) is easily seen to be a finite union of languages of the form $xu^*y$ ($x, u, y \in A^*$). Similarly, a language of polynomial density, also called *sparse*, can be shown to be a finite union of languages of the form $u_0^*v_1u_1^* \cdots v_nu_n^*$ where the $u_i$ and $v_j$ are in $A^*$. Together with $A^*$, the set of slender (resp. sparse) languages in $A^*$ forms a quotient-closed lattice of languages, for which defining profinite inequalities can be found in [22].

## 2.7 Deciding membership in an equationally defined class of languages

We are often interested in decision problems for families of regular languages: We say that a family $\mathcal{F}$ of regular languages over a finite alphabet $A$ is *decidable* if there is an algorithm that, given a regular language in $L \subseteq A^*$ as input, determines whether $L \in \mathcal{F}$. Here a regular language $L$ is 'given' by specifying a DFA that recognizes $L$, or some other formalism (*e.g.*, regular expression, logical formula) from which a DFA can be effectively computed. The problem arises, for example, if we are looking for a test of whether a given language is expressible in some logic for defining regular languages. (See Section 3.)

We can similarly define decidable families of finite monoids: Such a family $\mathcal{F}$ is decidable if there is an algorithm that, given the multiplication table for a finite monoid $M$, determines whether $M \in \mathcal{F}$. The definition extends in the obvious fashion to families of ordered monoids and stamps. For ordered monoids the input includes, in addition to the multiplication table of $M$, a representation of the graph of the partial order on $M$. For stamps $\varphi\colon A^* \to M$ we are also given the values $\varphi(a)$ for $a \in A$.

We will say that a variety $\mathcal{V}$ of languages is decidable if $A^*\mathcal{V}$ is decidable for every finite alphabet $A$. In this case the Eilenberg correspondence theorem gives a rather obvious connection between the two kinds of decidable families:

**Theorem 2.18.** *A (positive) variety (respectively, $\mathcal{C}$-variety) of languages is decidable if and only if the corresponding pseudovariety of (ordered) monoids (respectively, stamps) is decidable.*

*Proof.* We give the proof just for the case of ordinary varieties of languages and pseudovarieties of monoids; the argument is essentially the same for all the other variants. Let $\mathcal{V}$ be a variety of languages and **V** the corresponding pseudovariety of monoids. Suppose first that **V** is decidable. Let $\mathcal{A} = (Q, A, i, F)$ be a DFA recognizing a language $L \subseteq A^*$. From $\mathcal{A}$ we can effectively construct the multiplication table of $\mathrm{Synt}(L)$. We then apply the algorithm for **V** to decide whether $\mathrm{Synt}(L) \in \mathbf{V}$, and thus whether $L \in A^*\mathcal{V}$. Conversely, suppose $\mathcal{V}$ is decidable. Let $M$ be a finite monoid and choose a finite alphabet $A$ together with a surjective morphism $\varphi\colon A^* \to M$. (For example, we could choose $A = M$ and $\varphi$ the extension to $A^*$ of the identity map on $M$.) Then by Lemma 2.16 and Corollary 2.17, $M$ divides the direct product of the monoids $\mathrm{Synt}(\varphi^{-1}(m))$ for $m \in M$, and each of the $\mathrm{Synt}(\varphi^{-1}(m))$ in turn divides $M$. Thus $M \in \mathbf{V}$ if and only if each of the languages $\varphi^{-1}(m)$ is in $A^*\mathcal{V}$. Furthermore, from $\varphi$ we can construct a DFA $(M, A, 1, \{m\})$ recognizing $\varphi^{-1}(m)$, and thus decide whether each is in $A^*\mathcal{V}$. Thus **V** is decidable. □

Decision problems for varieties of regular languages can have arbitrarily large computational complexity, or indeed be undecidable. To see this, observe simply that if $P$ is *any* set of primes, then we can form the pseudovariety $\mathbf{G}_P$ of finite groups $G$ such that every prime divisor of $|G|$ is in $P$. Testing membership of a given prime $p$ in $P$ then reduces, in time polynomial in $p$, to testing membership in $\mathbf{G}_P$, so $\mathbf{G}_P$ is at least as complex as $P$.

On the other hand, Reiterman's theorem, which says varieties are defined by sets of profinite identities, suggests that we could determine membership in varieties simply by verifying whether identities hold in finite monoids. This is deceptive, since elements of $\widehat{X^*}$ do not generally have simple descriptions that make it possible to evaluate their images in finite monoids, and, further, the equational description of a pseudovariety might require inifinitely many profinite identities. We can nonetheless say something definitive about the complexity of the decision problems in the case where the equational definition consists of a finite set of profinite identities $\rho = \sigma$, where $\rho$ and $\sigma$ are $\omega$-terms in $\widehat{X^*}$: This means that $\rho$ and $\sigma$ are formed from elements of $X$ by successive application of concatenation and the operation $\tau \mapsto \tau^\omega$.

**Theorem 2.19.** *Let $\mathcal{V}$ be a variety of languages defined by a finite set of profinite identities of the form $\rho = \sigma$, where $\rho$ and $\sigma$ are $\omega$-terms, and let **V** be the corresponding pseudovariety of finite monoids. Then **V** is decidable by a logspace algorithm in the size of the input multiplication table, and $\mathcal{V}$ is decidable by a polynomial space algorithm in the size of the input automaton.*

*Proof.* We first consider testing membership of a monoid $M$ in **V**. Let $|M| = n$. The multiplication table of $M$ can be represented in $O(n^2 \log n)$ bits and each element of $M$

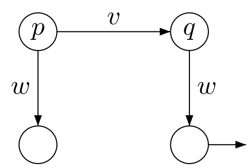
by $O(\log n)$ bits. We will show how to determine membership of $M$ in **V** using $k \cdot \log_2 n$ additional bits of workspace, where the constant $k$ is determined by the length of the longest $\omega$-term occurring in the defining profinite identities for **V**. To make the proof easier to follow, let us suppose we have an identity $((x^\omega y)^\omega z)^\omega = (xz)^\omega$. The algorithm loops through all triples $(x, y, z)$ of elements of $M$ and writes them in the workspace. It then uses $\log_2 n$ bits of additional workspace to compute $x^\omega$. This is done by repeatedly consulting the multiplication table, writing $x^2, x^3, \ldots$ in the same workspace, and after each write, consulting the multiplication table to check if the element is idempotent. We similarly compute $(x^\omega y)^\omega$, $((x^\omega y)^\omega z)^\omega$, and $(xyz)^\omega$. All in all, we used $7 \cdot \log_2 n$ bits of workspace. After all the values are computed, we compare the last two. The algorithm rejects if it finds a mismatch. If it finds none, it goes on to the next identity, and accepts if all the identities are tested with no mismatch.

We now turn to testing membership in $\mathcal{V}$. The algorithm we give is actually a nondeterministic polynomial space algorithm for nonmembership of a regular language in $A^*\mathcal{V}$. Since, by Savitch's Theorem ( [40], see, also Sipser [44]) nondeterministic polynomial space is equivalent to deterministic polynomial space, and the latter is closed under complement, this will be enough. Let us work with the same example identity we used in the first part of the proof. The algorithm begins by guessing words $x, y, z$ and computing the vectors

$$\begin{aligned}&(q_1x, \ldots, q_nx),\\ &(q_1y, \ldots, q_ny),\\ &(q_1z, \ldots, q_nz),\end{aligned}$$

where $\{q_1, \ldots, q_n\}$ is the set of states of the input DFA. Observe that the words $x, y, z$ themselves are not stored. Instead they are guessed letter by letter, and only the vectors of states are written in the workspace. This requires $O(n \log n)$ bits, where $n$ is the number of states of the DFA. Observe as well that once we have the vector $(q_1u, \ldots, q_nu)$ we can, with an additional $n \log_2 n$ bits, compute the vector $(q_1u^\omega, \ldots, q_nu^\omega)$, since we can write the vectors of the successive powers $(q_1u^k, \ldots, q_nu^k)$ reusing the same workspace, and then check after each write whether $qu^k = qu^{2k}$ for each state $q$. As a result we obtain the vectors $(q_1\hat\varphi(\rho), \ldots, q_n\hat\varphi(\rho))$, $(q_1\hat\varphi(\sigma), \ldots, q_n\hat\varphi(\sigma))$ for some morphism $\varphi\colon X^* \to A^*$. If these vectors turn out to be different, we accept. Thus this algorithm nondeterministically recognizes the complement of $A^*\mathcal{V}$, using $O(n \log n)$ space. □

The foregoing theorem illustrates a potentially large gap in complexity between testing membership in $\mathcal{V}$ from an input DFA and testing membership in the corresponding pseudovariety **V** from the multiplication table of a monoid. This is to be expected, since an automaton is in general exponentially more succinct than the multiplication table of its transition monoid. In some instances, however, it is possible to give efficient algorithms that begin with automata, using so-called 'forbidden pattern' characterizations of varieties. We illustrate this with a very simple example, using the ordered variety $\mathcal{J}^+$. Consider the following figure:




We say that a DFA $(Q, A, i, F)$ contains this pattern if there are states $q_1, q_2$ and words $u, v, w \in A^*$ such that $iu = q_1$, $q_2 = q_1 v$, $q_1 w \in F$, $q_2 w \notin F$. We say the DFA avoids the pattern if it does not contain it. It is easy to see that a DFA recognizing a language $L$ avoids this pattern if and only if whenever $uw \in L$, $uvw \in L$. Thus the languages in $A^*$ avoiding the pattern are exactly those that satisfy the inequality $x \leqslant 1$; that is, the language family $A^*\mathcal{J}^+$. We use this to prove the following:

**Theorem 2.20.** *There is an algorithm determining membership in $\mathcal{J}^+$ that runs in nondeterministic logspace in the size of an accepting DFA. (In particular, membership can be determined in polynomial time.)*

*Proof.* We nondeterministically guess letters to obtain an accessible state $q_1$, using $\log_2 n$ bits, where $n$ is the number of states in the automaton. We then further guess letters to obtain another state $q_2 = q_1 v$, written on another $\log_2 n$-bit field in the work space. Finally, we guess more letters, applying them to both components of the pair $(q_1, q_2)$ and arrive at at a state $(q_1 w, q_2 w)$. We accept if the first member of this pair of states is an accepting state of the DFA and the second is not. Thus we have a nondeterministic logspace algorithm for the regular languages outside of $\mathcal{J}^+$. But by the theorem of Immerman and Szelepcsenyi (see [26], [51], also [44]), nondeterministic logspace is closed under complement, so we have the desired result.

□

The same reasoning is used in many proofs showing that varieties of languages are decidable in nondeterministic logspace: find a forbidden pattern characterization of the variety using a fixed number of states. (For instance, Pin and Weil [37], Glasser and Schmitz [23].) While such results appear to bridge the complexity gap between polynomial-time algorithms that begin with a multiplication table and exponential-time algorithms that begin with an automaton, forbidden pattern arguments are not always available. In particular, we have the following result, which we cite without proof, from Cho and Huynh [17]:

**Theorem 2.21.** *Testing whether a regular language given by a DFA is aperiodic is PSPACE-complete.*

# 3 Connections with logic

In Section 1 we outlined, in an informal way, some of the logical apparatus for expressing properties of words over a finite alphabet. Here we give a more precise and general description. As before, variable symbols $x, y, x_1, x_2$, *etc.*, denote positions in a word. For

each $a \in A$ our logics have a unary predicate symbol $Q_a$, where $Q_a x$ is interpreted to mean 'the symbol in position $x$ is $a$.' We also have a binary predicate symbol $s$, where $s(x, y)$ is interpreted to mean 'position $y$ is the successor of position $x$'. We will usually use the alternative notation $y = x + 1$ for this.

We now consider *monadic second-order* formulas over this base of predicates. These are formulas built not merely by quantifying over individual positions, but also by quantifying over *sets* of positions, denoted by upper-case variable letters, and employing an additional relation symbol $x \in X$ between positions (first-order variables) and sets of positions (second-order variables).

For example, consider the monadic second order formula $\varphi$:

$$\exists x \exists y \exists X (Q_a x \wedge Q_b y \wedge x \in X \wedge y \in X \wedge \varphi_1 \wedge \varphi_2),$$

where $\varphi_1$ is

$$\neg \exists z (x = z + 1 \wedge z \in X) \wedge \neg \exists z (z = y + 1 \wedge z \in X),$$

and $\varphi_2$ is

$$\forall z (z \in X \to (y = z \vee \exists u (u \in X \wedge u = z + 1))).$$

The formula $\varphi$ is a *sentence*; that is, it has no free variables. Thus $\varphi$ defines a language $L_\varphi$ over $A = \{a, b\}$, namely the set of all words in which the formula is true. The sentence asserts the existence of positions $x$ and $y$ with letters $a$ and $b$ respectively, and of a set $X$ of positions that contains both $x$ and $y$, that contains the successor of each of its elements with the exception of $y$, and that contains no elements less than $x$. Thus $L_\varphi$ is the regular language $A^* a A^* b A^*$.

This example is an instance of the following important theorem, due to J. R. Büchi [16] (see [29, 49]).

**Theorem 3.1.** *A language $L \subseteq A^*$ is regular if and only if $L = L_\varphi$ for some sentence $\varphi$ of monadic second-order logic.*

We obtain subclasses of regular languages by restricting these second-order formulas in various ways. One obvious such restriction is to study first-order formulas: those formulas that use no second-order quantification. We denote this logic, as well as the family of regular languages that can be defined in it, by $\mathrm{FO}[+1]$. More generally, consider any $k$-ary relation $\alpha$ on the set of positions in a word that does not depend on the letters that appear in the word. Suppose further that $\alpha(x_1, \ldots, x_k)$ is definable by a formula of monadic second-order logic. Then we obtain a subclass of the regular languages by considering those languages definable by first-order sentences in which $\alpha$ is allowed as an atomic formula. We denote this class $\mathrm{FO}[\alpha]$, and similarly write $F[\alpha_1, \alpha_2, \ldots]$ when there are several such predicates. For example, the relation $x < y$ is definable in monadic second-order logic, by a formula much like the one used above to define the language $L = A^* a A^* b A^*$. Thus we obtain the logic and the language class $\mathrm{FO}[<]$. Of course, $L$ is definable in this logic, by the very simple sentence

$$\exists x \exists y (Q_a x \wedge Q_b y \wedge x < y).$$

We can extend the expressive power further, by adjoining, for $k > 1$, a binary predicate $\equiv_k$ that says two positions are equivalent modulo $k$. These predicates, too, are defin-

able in monadic second-order logic, and thus we obtain language classes $\mathrm{FO}[<, \equiv_k]$. We can further restrict these families by bounding the quantifier depth, or the alternation of existential and universal quantifiers, or the number of distinct variable symbols.

We are interested in understanding the expressive power of these logics, and determining exactly what languages can be defined in them. The critical insight is that, essentially, *(nearly) all these language classes are varieties.* In some instances we obtain ordered varieties, in others $\mathcal{C}$-varieties for a class $\mathcal{C}$ of morphisms, but in all cases we obtain families that, at least in principle, admit a characterizations in terms of the syntactic monoids and morphisms of the languages they contain.

### 3.1 Model-theoretic games

To see why this is so, we first describe an important tool for studying the expressive power of logics for words. Consider a first-order logic $\mathrm{FO}[\alpha_1, \ldots, \alpha_m]$. Look at a pair of words $w, w' \in A^*$ and suppose that on each word we have placed $k$ 'pebbles' labeled $x_1, \ldots, x_k$ for $w$, and $x'_1, \ldots, x'_k$ for $w'$. Each pebble is placed on a single position in its word, but two different pebbles can be on the same position. We denote the resulting pebbled words by $u = (w, x_1, \ldots, x_k)$ and $u' = (w, x'_1, \ldots, x'_k)$.

We will now describe a game $\mathcal{G}_r(u, u', \alpha_1, \ldots, \alpha_k)$ played on these two pebbled words. (This is called an *Ehrenfeucht Fraïssé game*.) The subscript $r$ denotes the number of rounds of the game. There are two players, traditionally called *Spoiler*, who plays first, and *Duplicator* who plays second. We define the rules of the game by induction on the number of rounds. In the 0-round game, the winner is already determined: If there is a relation $\alpha = \alpha_i$ of arity $p$, and pebbles $x_{i_1}, \ldots, x_{i_p}, x'_{i_1}, \ldots, x'_{i_p}$, such that

$$\alpha(x_{i_1}, \ldots, x_{i_p})$$

holds, and

$$\alpha(x'_{i_1}, \ldots, x'_{i_p})$$

does not, or vice-versa, then Spoiler wins the game. If there are pebbles $x_i$ and $x'_i$ such that the letter in position $x_i$ of $w$ is different from the letter in position $x'_i$ of $w'$, then Spoiler also wins the game. Otherwise, Duplicator wins. The idea is that Spoiler wins if the two pebbled words are different, and the difference must be witnessed by the atomic formulas applied to the pebbled positions.

Now let $r > 0$. In the $r$-round game $\mathcal{G}_r(u, u', \alpha_1, \ldots, \alpha_m)$, Spoiler makes a play by placing a new pebble $x_{k+1}$ in $u$ or $x'_{k+1}$ in $u'$. If Spoiler played in $u$ then Duplicator must respond with $x'_{k+1}$ in $u'$. Otherwise Duplicator responds with $x_{k+1}$ in $u$. The result is two new pebbled words $v, v'$. Spoiler and Duplicator proceed to play the game $\mathcal{G}_{r-1}(v, v', \alpha_1, \ldots, \alpha_m)$. Whoever wins this $(r-1)$-round game is the winner of the $r$-round game.

Ordinary words may be considered as special instances of pebbled words and thus we can consider the games $\mathcal{G}_r(w, w', \alpha_1, \ldots, \alpha_m)$, where $w, w' \in A^*$. The fundamental property of such games is given by the following theorem.

**Theorem 3.2.** *Let $w, w' \in A^*$, $r \geqslant 0$. The words $w$ and $w'$ satisfy the same sentences in $\mathrm{FO}[\alpha_1, \ldots, \alpha_m]$ of quantifier depth $r$ or less if and only if Duplicator has a winning*

*strategy in* $\mathcal{G}_r(w, w', \alpha_1, \ldots, \alpha_m)$.

See, for example, [29, 49].

Here is an example: Consider the two words $w = aab$ and $w' = aaab$. Spoiler has a winning strategy if $\mathcal{G}_2(w, w', <)$: First play pebble $x_1$ on the second $a$ of $w'$. If Duplicator replies on the first $a$ of $w$, Spoiler will play $x_2$ on the first $a$ of $w'$. If Duplicator instead replies on the second $a$ of $w$, then Spoiler plays $x_2$ on the third $a$ of $w'$. In either case, Duplicator has nowhere to play $x_2'$ in $w$ and win the game. By Theorem 3.2, there must be some sentence of quantifer depth 2 that distinguishes the two words. Indeed, $w'$ satisfies

$$\exists x(Q_a x \wedge \exists y(Q_a y \wedge x < y) \wedge \exists y(Q_a y \wedge y < x)),$$

while $w$ does not. On the other hand, Duplicator has a winning strategy in the two-round game in $aaaab, aaab$.

What does this have to do with varieties? We will use games to show that logically-defined language classes satisfy the closure properties that define varieties. Look, for example, at the family of languages defined by $\mathrm{FO}[<]$ sentences of quantifier depth no more than $d$, where $d \geqslant 0$. We will denote both this language family and the underlying logic by $\mathrm{FO}_d[<]$.

**Theorem 3.3.** $\mathrm{FO}_d[<]$ *is a variety of languages.*

*Proof.* Since we have to discuss languages over different alphabets, let us denote by $A^* \,\mathrm{FO}_d[<]$ the languages over $A^*$ that belong to this family. Obviously $A^* \,\mathrm{FO}_d[<]$ is closed under boolean operations, so we must verify closure under quotients and inverse images of morphisms. Let us write $w \sim_{d,A} w'$ to mean that $w, w' \in A^*$ satisfy all the same sentences of $\mathrm{FO}_d[<]$. Then $\sim_d$ is an equivalence relation of finite index on $A^*$, and every language of $A^* \,\mathrm{FO}_d[<]$ is a union of $\sim_{d,A}$-classes. We claim that if $w \sim_{d,A} w'$ and $a \in A$, then both $aw \sim_{d,A} aw'$, and $wa \sim_{d,A} w'a$, and that further, if $\varphi\colon A^* \to B^*$ is a morphism, then $\varphi(w) \sim_{d,B} \varphi(w')$.

To see that this claim implies the result, suppose $L \in A^* \,\mathrm{FO}_d[<]$ but $a^{-1}L \notin A^* \,\mathrm{FO}_d[<]$. Then there exist $w, w' \in A^*$ with $w \in a^{-1}L$, $w' \notin a^{-1}L$, and $w \sim_{d,A} w'$. But then $wa \in L$, $w'a \notin L$, and $wa \sim_{d,A} w'a$, contradicting $L \in A^* \,\mathrm{FO}_d[<]$. By the same reasoning we deduce closure under right quotients and under inverse images of morphisms.

To prove the claim, note that by Theorem 3.2, $w \sim_{d,A} w'$ if and only if Duplicator has a winning strategy in $\mathcal{G}_d(w, w', <)$. So we must show that such a winning strategy implies the existence of winning strategies for Duplicator in $\mathcal{G}_d(aw, aw', <)$, $\mathcal{G}_d(wa, w'a, <)$, and $\mathcal{G}_d(\varphi(w), \varphi(w'), <)$. For $\mathcal{G}_d(wa, w'a, <)$, the strategy is this: Whenever Spoiler plays on the last letter of either $wa$ or $w'a$, Duplicator responds by playing on the last letter of the other word; otherwise Duplicator responds according to the winning strategy in $(w, w')$. The reasoning is identical for $\mathcal{G}_d(aw, aw', <)$. For $\mathcal{G}_d(\varphi(w), \varphi(w'), <)$, suppose $w = a_1 \cdots a_r$, $w' = a_1' \cdots a_s'$, and let $v_i = \varphi(a_i)$, $v_i' = \varphi(a_i')$. Duplicator's strategy is to keep track of a separate game in $w, w'$ to calculate the responses in $\varphi(w), \varphi(w')$. If Spoiler plays on the $j^{th}$ symbol of $v_i$, then Duplicator calculates the response, according to the original strategy, to a move by Spoiler on $a_i$. Let us say this response is on $a_k'$. Observe that $a_i = a_k'$, and thus $v_i = v_k'$, so Duplicator can reply on the $j^{th}$ symbol of $v_k'$.

In other words, Duplicator pulls the Spoiler's plays back to $(w, w')$, applies the original winning strategy, and pushes the result forward to $(\varphi(w), \varphi(w'))$. It is easy to see that this strategy wins for Duplicator. $\square$

This same reasoning can be adapted to a large number of different situations. Consider, for example, the logics $\mathrm{FO}_d[+1]$. The strategy-copying argument no longer works to give Duplicator a winning strategy in $\mathcal{G}_d(\varphi(w), \varphi(w'), +1)$, because $\varphi$ may map a letter to the empty word, and thus we might end up with two pebbles on adjacent positions in $\varphi(w)$, but find the corresponding pebbles on non-adjacent positions of $\varphi(w')$. But the argument *does* work for non-erasing morphisms, and thus each $\mathrm{FO}_d[+1]$, as well as the union $\mathrm{FO}[+1]$, is a $\mathcal{C}_{ne}$-variety. Similarly, suppose we augment the logic $\mathrm{FO}[<]$ by adjoining the predicate $x \equiv_q y$ for equivalence modulo $q$. We now find that the strategy-copying argument works as long as all $\varphi(a)$ for $a \in A$ have the same length $m$, as $i \equiv_q j$ implies $mi \equiv_q mj$. Thus each $\mathrm{FO}_d[<, \equiv_q]$ is a $\mathcal{C}_{lm}$-variety of languages.

This reasoning is amenable to further adaptations, by altering the rules of the games: We obtain a game characterization of languages defined by formulas that use no more than $p$ distinct variables by allowing only $p$ pebbles, regardless of the number of rounds. Once all the pebbles have been placed, the Spoiler may pick up a pebble and move it to a new position; the Duplicator must pick up the corresponding pebble and move it in the same direction. We obtain a game characterization of the languages defined by boolean combinations of $\Sigma_k$ sentences, with quantifier block size bounded by $d$, by considering $k$-round games in which each player is permitted to place $d$ pebbles at a time. We can turn this into a game characterization of the languages defined by $\Sigma_k$-sentences themselves by requiring Spoiler to play in $w$ in the first round, in $w'$ in the second round, *etc.* Duplicator then has a winning strategy in the game in $w, w'$ if and only if every $\Sigma_k$-sentence, with quantifier block size no more than $d$, that $w$ satisfies is also satisfied by $w'$. We can use this to conclude that $\Sigma_k[<]$ is an ordered variety of languages. In all instances, we find that some variant of Eilenberg's Theorem applies, and extract the same conclusion: A logical characterization of the language class implies the existence of an algebraic characterization.

Care must be taken not to extrapolate this *too* far. For example, the strategy-copying argument fails in the case of $\Sigma_1[+1]$: Let $w = abab$, $w' = baba$. Then $w, w'$ satisfy the same $\Sigma_1[+1]$-sentences of block size 2, but $wa$ and $w'a$ do not, since $w'a$ contains two consecutive occurrences of $a$.

## 3.2 Explicit characterization of logically defined classes

While the foregoing arguments tell us that logically defined language classes form varieties, they do not provide explicit algebraic characterizations. There are, in fact, a number of different methods for connecting the structure of defining sentences to algebraic properties, and many results giving explicit characterizations of the language varieties defined by various logics. (See, for instance Straubing [49].) Here we give just a taste of these techniques and results with what is perhaps the most famous, and certainly the first, result in this area, the theorem of McNaughton and Papert [30] giving the equivalence of first-order logic and aperiodic monoids:

**Theorem 3.4.** *A language $L$ belongs to* $\mathrm{FO}[<]$ *if and only if* $\mathrm{Synt}(L)$ *is aperiodic.*

We will only prove one direction of this theorem, namely that first-order definability implies aperiodicity. We claim that if $u \in A^*$, then $u^{2^d-1} \sim_{d,A} u^{2^d}$. This is proved by induction on $d$. For $d = 0$, there is nothing to prove, since all words are equivalent modulo $\sim_{0,A}$ . Suppose then that $d > 0$. We will show that Duplicator has a winning strategy in $\mathcal{G}_d(u^{2^d-1}, u^{2^d}, <)$. Suppose Spoiler plays $x_1$ in $u^{2^d-1}$.

$$u^{2^d-1} = u^r vav' u^s,$$

where the pebble is played on the position indicated by the letter $a$, $u = vav'$, and $r+s = 2^d - 2$. It follows that either $r \geqslant 2^{d-1} - 1$ or $s \geqslant 2^{d-1} - 1$. Suppose the former (the proof is the same in either case). Then we can write

$$u^{2^d} = u^{r+1} vav' u^s.$$

Duplicator places the pebble $x_1'$ on the indicated $a$. Now play proceeds as follows: By the inductive hypothesis, Duplicator has a winning strategy in $\mathcal{G}_{d-1}(u^{2^{d-1}-1}, u^{2^{d-1}})$. Thus, by the argument given in the proof of Theorem 3.3, Duplicator has a winning strategy in $\mathcal{G}_{d-1}(u^r v, u^{r+1} v)$. Duplicator will follow this strategy whenever Spoiler plays to the left of $x_1$ or $x_1'$, and simply copy Spoiler's move in $av'u^s$ whenever the play is at or to the right of $x_1$ or $x_1'$. This proves the claim. It follows that if $L$ is first-order definable, then $\mathrm{Synt}(L)$ satisfies the $x^m = x^{m+1}$ for sufficiently large $m$, and is thus aperiodic.

We omit the proof of the converse, that if $\mathrm{Synt}(L)$ is aperiodic, then $L$ is in $\mathrm{FO}[<]$. Most of the published proofs of this theorem rely on some decomposition theory for finite semigroups, either the Krohn-Rhodes decomposition, or the ideal structure of semigroups. Most proofs also show first that every language recognized by an aperiodic monoid is a star-free language. We will define star-free languages in Section 4.2, and show that they are equivalent to first-order definable languages. Pin [32] gives a relatively streamlined proof using the ideal decomposition theory. Straubing [49] uses the Krohn-Rhodes decomposition to obtain a first-order sentence directly. Wilke [55] gives a proof that is remarkable for its absence of hard semigroup theory, and that produces a formula of temporal logic directly from an automaton with an aperiodic transition monoid. □

We can use Theorem 3.4 to deduce a claim we made earlier, giving an explicit characterization of the $\mathcal{C}_{lm}$-pseudovariety **QA**:

**Theorem 3.5.** *$L$ belongs to* $\mathrm{FO}[<, \equiv_m]$ *for some $m > 1$ if and only if the syntactic morphism of $L$ is in* $\mathbf{QA}$.

We merely sketch the argument: Suppose $u \in A^+$ with $|u|$ divisible by $m$. Let $d > 0$. Then by precisely the same argument as we gave in the proof of Theorem 3.4, Duplicator has a winning strategy in $\mathcal{G}_d(u^r, u^{r+1}, <, \equiv_m)$ as long as $r$ is sufficiently large compared to $d$. This is enough to show that if $L$ is definable by a sentence of $\mathrm{FO}[<, \equiv_m]$, then the stable semigroup of $\eta_L$ is aperiodic. For the converse, we consider a language $L$ with $\eta_L$ in **QA**. Let $\eta_L(A^t)$ be the stable semigroup. If we treat $B = A^t$ as a finite alphabet, we can use Theorem 3.4 to obtain a first-order sentence, with respect to $B$, defining the sets of words of length divisible by $t$ that are recognized by $\eta_L$, and then translate this to a first-order sentence over $A$ by means of the predicate $\equiv_t$ .

**Other logical formalisms** By and large, we have confined our discussion of logic to the use of first-order quantification. But there are other formalisms studied in the literature, which also give rise to varieties. We mention in passing two of these: Formulas with modular quantifiers, which were introduced by Straubing, Thérien and Thomas [50] and studied extensively in [49], and *temporal* formulas, which play an important role in computer-aided verification. An algebraic treatment of temporal logic, and its connection to varieties of languages, is due to Thérien and Wilke [53, 54] and Wilke [55, 56].

# 4 Operations on classes of languages

The idea developed in this section is that certain operations on classes of languages translate to operations on the corresponding sets of profinite identities, or on the corresponding classes of syntactic objects (syntactic monoids or semigroups, ordered or not, etc). This translation, when it can be made explicit, may provide decomposition results, or membership decision results for complex classes of languages.

## 4.1 Boolean operations

If for each $i \in I$, $\mathcal{V}_i$ is a class of regular languages, the intersection $\mathcal{W} = \bigcap_{i\in I} \mathcal{V}_i$ is the class given by $A^*\mathcal{W} = \bigcap_{i\in I} A^*\mathcal{V}_i$ for each alphabet $A$. The different classes of families of languages considered so far (lattices or boolean algebras of languages of some fixed $A^*$, positive $\mathcal{C}$-varieties) are easily seen to be closed under (arbitrary) intersection.

The following statement essentially follows from the definition of the satisfaction of profinite equations.

**Proposition 4.1.** *Let $I$ be a set and for each $i \in I$, let $E_i$ be a set of profinite equations on an alphabet $A$. Then $\bigcap_{i\in I} \mathcal{L}(E_i) = \mathcal{L}(\bigcup_{i\in I} E_i)$.*

*In particular, if for each $i \in I$ $\mathcal{V}_i$ is a class of regular languages that is $\mathcal{C}$-defined by a set of profinite (ordered) $\mathcal{C}$-identities $E_i$, then $\bigcap_{i\in I} \mathcal{V}_i$ is $\mathcal{C}$-defined by $\bigcup_{i\in I} E_i$.*

The fact that an arbitrary intersection of lattices of regular languages (resp. (positive) $\mathcal{C}$-varieties) is again a lattice of regular languages (resp. a (positive) $\mathcal{C}$-variety) has the following consequence: for each set $V$ of regular languages in $A^*$ (resp. every class $\mathcal{V}$ of regular languages) there exists a least lattice (resp. a least (positive) $\mathcal{C}$-variety) containing it, which is said to be *generated by* $V$ (resp. $\mathcal{V}$).

The union of two lattices of languages in $A^*$ is not a lattice in general. The relevant operation is the *join*: the join of two lattices of regular languages in $A^*$ (resp. classes of regular languages) is defined to be the lattice generated by their union.

Describing the profinite equations or identities defining a join is difficult. In fact, Albert, Baldinger and Rhodes exhibit [1] a finite set $\Sigma$ of computable profinite identities, such that the join of the pseudovariety $[\![\Sigma]\!]$ with the pseudovariety **Com** $= [\![xy = yx]\!]$ of commutative monoids, is not decidable (see also [7]).

Some joins were computed early, based on the structural theory of monoids. This is the case for instance of $\mathbf{J}_1 \vee \mathbf{G}$, which is characterized as the class of finite monoids which

are unions of groups and in which idempotents commute (see [25]). This translates as

$$\mathbf{J}_1 \vee \mathbf{G} = [\![x^{\omega+1} = x,\ x^\omega y^\omega = y^\omega x^\omega]\!].$$

Other joins resisted computation until the advent of profinite methods, such as the joins $\mathbf{R} \vee \mathbf{L}$ (Almeida and Azevedo [4]) and $\mathbf{G} \vee \mathbf{Com}$ (Almeida [2]). The case of $\mathbf{J} \vee \mathbf{G}$ is interesting, since this join is decidable but is not defined by a finite set of profinite identities (Almeida, Azevedo and Zeitoun [5], Steinberg [45, 46], Trotter and Volkov [58]).

**Example 4.1.** The following simple examples will be useful in the sequel. Let $\mathbf{I} = [\![x = y]\!]$ be the trivial pseudovariety of monoids (which consists only of the 1-element monoid). Let also $\mathbf{K}$ and $\mathbf{D}$ be the pseudovariety of semigroups $\mathbf{K} = [\![x^\omega y = x^\omega]\!]$ and $\mathbf{D} = [\![yx^\omega = x^\omega]\!]$. The elements of $\mathbf{K}$ are the finite semigroups in which idempotents act as zeroes on the left. Dually, in the semigroups of $\mathbf{D}$, idempotents act like zeroes on the right. If $\mathbf{V}$ is any pseudovariety of monoids, we let $L\mathbf{V}$ be the class of finite semigroups $S$ such that $eSe \in \mathbf{V}$ for each idempotent $e$ of $S$. It is easily verified that $L\mathbf{V}$ is a pseudovariety of semigroups, and that it is decidable if and only if $L\mathbf{V}$ is.

It is easily verified that the semigroups that are both in $\mathbf{K}$ and in $\mathbf{D}$ are exactly the semigroups with a single idempotent, which is a zero (these semigroups are called *nilpotent*). Interestingly, the join $\mathbf{K} \vee \mathbf{D}$ is equal to $L\mathbf{I} = [\![x^\omega y x^\omega = x^\omega]\!]$.

## 4.2 Closure operations and Mal'cev products

An early closure result is Schützenberger's theorem on star-free languages. The set of *star-free languages* on alphabet $A$ is the least boolean algebra containing the letters (and the empty set), which is closed under concatenation. For instance, $aA^*$ is star-free, since it is equal to $a\emptyset^c$. A non-trivial question is that of decidability: given a regular language $L$, can we decide whether it is star-free? As it turns out, $(ab)^*$ is star-free (its complement is the set of all words with two consecutive $a$'s or two consecutive $b$'s, or that start with $b$ or end with $a$) but $(aa)^*$ is not...

The solution to this problem was given by Schützenberger [41] with the following theorem.

**Theorem 4.2.** *The class of star-free languages forms a variety of languages, corresponding to the pseudovariety* $\mathbf{Ap}$ *of aperiodic monoids. In particular, this class is decidable.*

In view of Theorem 3.4, this is equivalent to the following statement.

**Theorem 4.3.** *A language is star-free if and only if it is* $\mathrm{FO}[<]$*-definable.*

*Proof.* We prove Theorem 4.3 using game-theoretic methods, as in Section 3. Let us first show that a $\mathrm{FO}[<]$-definable language is star-free. It is sufficient to show, by induction on $k$, that for all $w \in A^*$ and $k \geqslant 0$, $[w]_k$ is star-free. The case $k = 0$ is trivial, since $[w]_0 = A^*$ for all $w \in A^*$. To prove the general case, we will establish the equality

$$[w]_{k+1} = \bigcap [x]_k a [x']_k \setminus \bigcup [y]_k b [y']_k,$$

where the intersection is over all factorizations $w = xax'$ with $x, x' \in A^*$ and $a \in A$, and the union is over all triples $([y]_k, b, [y']_k)$, where $b \in A$ and $w \notin [y]_k b [y']_k$. By induction, the $\sim_k$-classes are star-free languages, so the equality above implies that the $\sim_{k+1}$-classes are star-free as well.

To prove the equality, note that the inclusion from left to right is trivial, so we need only show that if $w' \in A^*$ is in the set on the right-hand side, then $w \sim_{k+1} w'$. So we will show that Duplicator has a winning strategy in the $(k+1)$-round game in the two words. Observe that inclusion of $w'$ in the right-hand side means that $w, w'$ have precisely the same set of factorizations with respect to $\sim_k$, in the sense that for every factorization $xax'$ of one word, with $a \in A$, there exists a corresponding factorization $yay'$ of the other word with $x \sim_k x'$, $y \sim_k y'$. Thus if Spoiler plays on a position in one of the words, inducing a factorization $xax'$ of the word, Duplicator can play on the corresponding position of the other. Duplicator can now correctly reply in the remaining $k$ rounds of the game by using his winning strategy in the games in $(x, y)$ and $(x', y')$.

Conversely, let us show that every star-free language is $\mathrm{FO}[<]$-definable. In view of the definition of star-free languages, we need to show, first, that $A^*$ and every language of the form $\{a\}$ ($a \in A$) is $\mathrm{FO}[<]$-definable; and second that if $K$ and $L$ are $\mathrm{FO}[<]$-definable, then so are the boolean combinations of $K$ and $L$, and so is $KL$. The only non-trivial point concerns the concatenation product, and the problem easily reduces to showing that $KaL$ ($a \in A$) is $\mathrm{FO}[<]$-definable.

Let us assume that $K$ and $L$ are defined by formulas of quantifier-depth $k$. Let $w \in KaL$, say, $w = uav$ with $u \in K$ and $v \in L$. We want to show that if $w \sim_{k+1} w'$ — that is, Duplicator has a winning strategy for$\mathcal{G}_{k+1}(w, w')$ —, then $w' \in KaL$. Let Spoiler put a pebble on the letter $a$ in $w$ witnessing the factorization $w = uav$, then Duplicator's strategy has her put a pebble on a letter $a$ in $w'$, determining a factorization $w' = u'av'$. We claim that Duplicator wins the $k$-round game in $u$ and $u'$: indeed, such a game can be seen as the 2nd, …, $(k+1)$-st moves in a game in $w = uav$ and $w' = u'av'$. Therefore $u \sim_k u'$ and hence $u' \in K$. Similarly $v' \in L$: thus $w' \in KaL$. □

A natural extension of the question answered by Schützenberger's theorem is the following: can we characterize the varieties of languages which are closed under concatenation product? and if $\mathcal{V}$ is a variety of languages, can we describe the least variety containing $\mathcal{V}$ and closed under concatenation product? Both problems were solved by Straubing [47]. In order to state his result, we need to introduce an operation on pseudovarieties.

Let **V** be a pseudovariety of monoids and let **W** be a pseudovariety of semigroups (resp. ordered semigroups). We consider the class of all finite monoids (resp. ordered monoids) $M$ for which there exists a morphism (un-ordered) $\varphi\colon M \to N$ such that $N \in$ **V** and $\varphi^{-1}(e) \in$ **W** for each idempotent element $e$ of $N$. This class is not a pseudovariety in general, but it is elementary to verify that the quotients (resp. ordered quotients) of its elements form a pseudovariety of monoids (resp. ordered monoids), called the *Mal'cev product* of **V** by **W**, and denoted **W** Ⓜ **V**.

**Theorem 4.4.** *Let $\mathcal{V}$ be a variety of languages and let **V** be the corresponding pseudovariety of monoids. If $\mathcal{W}$ is the least variety of languages containing $\mathcal{V}$ and closed under concatenation product, then the corresponding pseudovariety of monoids is **Ap** Ⓜ **V**.*

Schützenberger's theorem above is the particular case of Theorem 4.4 when $\mathcal{V}$ is the

trivial variety of languages.

Interestingly, the Mal'cev product is also useful to characterize the closure of a variety of languages under other types of products. For technical reasons, the definition of these products involves intermediate, marker letters: If $K$ and $L$ are languages in $A^*$, and if $a \in A$, we say that the product $KaL$ is *deterministic* if each word $u \in KaL$ has a unique prefix in $Ka$. Co-deterministic products are defined dually: the product $KaL$ is *co-deterministic* if each word $u \in KaL$ has a unique suffix in $aL$. Another important modality of product is the following: a product $L_0a_1L_1 \cdots a_kL_k$ is *unambiguous* if every word $u$ in this language admits a unique decomposition in the form $u = u_0a_1u_1 \cdots a_ku_k$ with each $u_i \in L_i$. Deterministic and co-deterministic products are particular cases of unambiguous products.

It is natural to extend these operations to classes of languages. Given a class of languages $\mathcal{V}$, we denote by $\mathrm{Det}\,\mathcal{V}$ the class of languages such that, for each alphabet $A$, $A^*\,\mathrm{Det}\,\mathcal{V}$ is the set of all boolean combinations of languages of $A^*\mathcal{V}$ and of deterministic products of these languages. $\mathrm{Det}\,\mathcal{V}$ is called the *deterministic closure* of $\mathcal{V}$. The *co-deterministic closure* $\mathrm{coDet}\,\mathcal{V}$ and the *unambiguous closure* $\mathrm{UPol}\,\mathcal{V}$ are defined similarly. Schützenberger [42, 32] characterized algebraically these operations for varieties of languages.

**Theorem 4.5.** *Let $\mathcal{V}$ be a variety of languages and let $\mathbf{V}$ be the corresponding pseudovariety of monoids. Then $\mathrm{Det}\,\mathcal{V}$, $\mathrm{coDet}\,\mathcal{V}$ and $\mathrm{UPol}\,\mathcal{V}$ are varieties of languages, and the the corresponding pseudovarieties of monoids are $\mathbf{K} \textcircled{m} \mathbf{V}$, $\mathbf{D} \textcircled{m} \mathbf{V}$ and $L\mathbf{I} \textcircled{m} \mathbf{V}$, respectively.*

**Example 4.2.** Consider the variety of languages $\mathcal{J}_1$, described in Sections 1.1 and 2.6.1: for each alphabet $A$, $A^*\mathcal{J}_1$ is the boolean algebra generated by the languages of the form $B^*$, with $B \subseteq A$. It is elementary to verify that $A^*\,\mathrm{Det}\,\mathcal{J}_1$ is the boolean algebra generated by the products of the form $A_0^*a_1A_1^* \cdots a_kA_k^*$, such that for each $0 < i \leqslant k$, $a_i \notin A_{i-1}$. Theorem 4.5 tells us that $\mathrm{Det}\,\mathcal{J}_1$ forms a variety of languages, and that the corresponding pseudovariety of monoids is $\mathbf{K} \textcircled{m} \mathbf{J}_1$.

Semigroup theory helps us characterize this pseudovariety. $\mathbf{K} \textcircled{m} \mathbf{J}_1$ is the class $\mathbf{R}$ of all so-called $\mathcal{R}$-trivial finite monoids, that is, the monoids $M$ in which principal right ideals have a single generator: $sM = tM$ implies $s = t$. In addition, one can show that $\mathbf{R} = [\![(xy)^\omega x = (xy)^\omega]\!]$. This induces immediately the decidability of $\mathrm{Det}\,\mathcal{J}_1$.

A dual result characterizes $\mathbf{D} \textcircled{m} \mathbf{J}_1$, the pseudovariety associated with $\mathrm{coDet}\,\mathcal{J}_1$, as the class $\mathbf{L}$ of $\mathcal{L}$-trivial finite monoids. It is interesting to note that $\mathbf{R} \cap \mathbf{L} = \mathbf{J}$. The variety of piecewise testable languages discussed in Section 1.2 is therefore the class of languages that can be described simultaneously as boolean combinations of deterministic and of co-deterministic products of the form $A_0^*a_1A_1^* \cdots a_kA_k^*$ with each $A_i$ a subset of $A$.

Similarly, Theorem 4.5 shows that the pseudovariety of monoids corresponding to $\mathrm{UPol}\,\mathcal{J}_1$ is is $L\mathbf{I} \textcircled{m} \mathbf{J}_1$. Again, one can show that this pseudovariety is the class of finite monoids in which every regular element is idempotent, usually denoted by **DA**, and equal to $[\![(xyz)^\omega z(xyz)^\omega = (xyz)^\omega]\!]$. It follows, here too, that $\mathrm{UPol}\,\mathcal{J}_1$ is decidable. Let us note in addition that it coincides with the class of languages that can be defined by $\mathrm{FO}[<]$ sentences that use at most two variable symbols. (See [52].)

The following result is of the same nature as Theorems 4.4 and 4.5 but it involves a

positive variety of languages, and the corresponding pseudovariety of ordered monoids. If $\mathcal{L}$ is a set of regular languages in $A^*$, we denote by $\operatorname{Pol}\mathcal{L}$ (the *polynomial closure* of $\mathcal{L}$), the lattice generated by the languages of the form $L_0a_1L_1\cdots a_kL_k$, with $L_i \in \mathcal{L}$ and $a_i \in A$ for each $i$. If $\mathcal{V}$ is a class of regular languages, then $\operatorname{Pol}\mathcal{V}$ is the class such that, for each alphabet $A$, $A^*\operatorname{Pol}\mathcal{V} = \operatorname{Pol}(A^*\mathcal{V})$. Then the following result holds, see [36].

**Theorem 4.6.** *Let $\mathcal{V}$ be a variety of languages. Then $\operatorname{Pol}\mathcal{V}$ is a positive variety of languages, and the the corresponding pseudovariety of ordered monoids is $[\![x^\omega yx^\omega \leqslant x^\omega]\!] \,ⓜ\, \mathbf{V}$.*

In general, the results reported above do not provide explicit decision algorithms, even if **V** is decidable (see [7]). However, the structural theory of semigroups yields some such results. In particular, we can use a result by Krohn, Rhodes and Tilson [28] to show that if $\mathcal{V}$ is decidable, then so are $\operatorname{Det}\mathcal{V}$, $\operatorname{coDet}\mathcal{V}$ and $\operatorname{UPol}\mathcal{V}$ (generalizing the specific instances discussed in Example 4.2).

It is not known whether **Ap** ⓜ **V** is decidable whenever $\mathcal{V}$ is. A positive solution to this problem would imply a positive solution to an open instance of the complexity problem, which we discuss below in 4.3.

Topological methods also [36] provide sets of profinite identities describing Mal'cev products. In the cases of interest for us, it yields the following statement.

**Proposition 4.7.** *Let $\mathcal{V}$ be a variety of languages. Then the least variety containing $\mathcal{V}$ and closed under concatenation is defined by the set of profinite identities of the form $x^{\omega+1} = x^\omega$, where $x \in \widehat{X^*}$ and **V** satisfies $x = x^2$.*

*Similar statements hold for $\operatorname{Det}\mathcal{V}$ (respectively, $\operatorname{coDet}\mathcal{V}$, $\operatorname{UPol}\mathcal{V}$ and $\operatorname{Pol}\mathcal{V}$), replacing the profinite identity $x^{\omega+1} = x^\omega$ by $x^\omega y = x^\omega$ (respectively, $yx^\omega = x^\omega$, $x^\omega yx^\omega = x^\omega$ and $x^\omega yx^\omega \leqslant x^\omega$), where $x, y \in \widehat{X^*}$ and **V** satisfies $x = x^2 = y$.*

These results were extended to $\mathcal{C}$-varieties, and in the case of $\operatorname{Pol}\mathbf{V}$, to lattices of regular languages closed under quotients [34, 15]. In practice, the resulting sets of profinite identities are infinite and incomputable. However, in a number of situations, one can extract from these sets more manageable, yet sufficient subsets, yielding decision algorithms.

**Example 4.3.** Branco and Pin [15] use Proposition 4.7—applied to the lattice of slender languages (see Section 2.6.8)— to prove the decidability of the lattice generated by the languages of the form $L_0a_1L_1\cdots a_kL_k$ where the $L_i$ are either $A^*$ or of the form $u^*$ for some $u \in A^*$.

## 4.3 Product operations and semidirect products

We now consider products of the form $LaA^*$, where $L$ is a language and $a \in A$: $LaA^*$ is the language of all words with a prefix in $La$. Given a monoid $M$ accepting $L$, one can construct a monoid accepting $LaA^*$ using the operation of semidirect product.

In general, let $S$ and $T$ be monoids. A *left action* of $T$ on $S$ is a mapping $\lambda\colon T \times S \to S$, written $(t, s) \mapsto t \cdot s$, such that for each $t$, the map $\lambda_t\colon s \mapsto t \cdot s$ is an endomorphism of

$S$, and such that the map $t \mapsto \lambda_t$ is a morphism from $T$ to the monoid of endomorphisms of $S$. Once such an action $\lambda$ is given, the *semidirect product* $S *_\lambda T$ (we usually write $S * T$) is the monoid of all pairs $(s,t) \in S \times T$, with product

$$(s,t)(s',t') = (s\,\lambda(t,s'),\ tt').$$

**Lemma 4.8.** *If* $\varphi\colon A^* \to T$ *accepts the language* $L$*, then* $U_1^T * T$ *accepts* $LaA^*$.

*Proof.* We consider the action $\lambda$ of $T$ on $U_1^T$ given by $\lambda(t,(s_x)_{x\in T}) = (s'_x)_{x\in T}$, with $s'_x = s_{xt}$. Let then $\psi\colon A^* \to U_1^T * T$ be given by , for each $b \in A$,

$$\psi(b) = \Big((s_x^{(b)})_{x\in T},\ \varphi(b)\Big) \qquad \text{with}$$
$$s_x^{(b)} = \begin{cases} 0 & \text{if } x \in \varphi(L) \text{ and } b = a,\\ 1 & \text{otherwise.}\end{cases}$$

Using the definition of the product in $U_1^T * T$, we find that

$$\begin{aligned}\psi(a_1\cdots a_n) &= ((r_x)_{x\in T},\ \varphi(a_1\cdots a_n)) \qquad \text{with}\\ r_x &= s_x^{(a_1)}\ s_{x\varphi(a_1)}^{(a_2)}\ \cdots\ s_{x\varphi(a_1\cdots a_{n-1})}^{(a_n)}\\ &= \begin{cases} 0 & \text{if for some } 1 \leqslant i \leqslant n,\ x\varphi(a_1\cdots a_{i-1}) \in \varphi(L) \text{ and } a_i = a,\\ 1 & \text{otherwise.}\end{cases}\end{aligned}$$

In particular, we observe that $a_1\cdots a_n \in LaA^*$ if and only if $r_1 = 0$. □

**Remark 4.9.** Observe that the construction of the semidirect product $U_1^T * T$ given above does not use anything special about $U_1$, and thus can be applied to any pair of monoids $U$ and $T$. This is called the *wreath product* $U \circ T$. The wreath product is closely related to the semidirect product, in the sense that first, it is, of course, a semidirect product with $T$ of a member of the pseudovariety generated by $U$, and, second, every semidirect product $U * T$ embeds in $U \circ T$. The wreath product, in a sense that can be made precise, captures the notion of serial composition of automata [18]. As a consequence it is frequently used, exactly as in the proof of Lemma 4.8 above to prove decomposition results.

The operation of semidirect product is naturally extended to pseudovarieties: if **V** and **W** are pseudovarieties, we let **V** $*$ **W** be the pseudovariety generated by the semidirect products $S * T$ with $S \in$ **V** and $T \in$ **W**. Then we have the following theorem.

**Theorem 4.10.** *Let* $\mathcal{V}$ *be a variety of languages and for each alphabet* $A$*, let* $A^*\mathcal{W}$ *bethe boolean algebra generated by the languages of* $A^*\mathcal{V}$ *and the languages of the form* $LaA^*$ *with* $L \in A^*\mathcal{V}$*. Then the class of languages* $\mathcal{W}$ *is a variety and the corresponding pseudovariety of monoids is* $\mathbf{J}_1 * \mathbf{V}$*.*

*Proof.* Since $U_1 \in \mathbf{J}_1$, Lemma 4.8 shows that every language in $A^*\mathcal{W}$ is accepted by a monoid in $\mathbf{J}_1 * \mathbf{V}$. The proof of the converse is a particular case of the more general *wreath product principle* (Straubing [48]). Let $\varphi$ be a morphism $\varphi\colon A^* \to S * T$ and for

each $a \in A$, let $\varphi(a) = (s_a, t_a)$. Let $\psi\colon A^* \to T$ be the morphism given by $\psi(a) = t_a$. Let also $B = T \times A$ and let $\sigma\colon A^* \to B^*$ be the map

$$\sigma(a_1 \cdots a_n) = (1, a_1)\ (\psi(a_1), a_2)\ \cdots\ (\psi(a_1 \cdots a_{n-1}), a_n).$$

Note that $\sigma$ is a so-called sequential function [11, 39], not a morphism. We observe however that, if $\chi\colon B^* \to S$ is the morphism given by $\chi(t, a) = t \cdot s_a$, then

$$\varphi(a_1 \cdots a_n) = (\chi\sigma(a_1 \cdots a_n),\ \psi(a_1 \cdots a_n))\,.$$

It follows that if $(s, t) \in S * T$, then $\varphi^{-1}(s, t) = \psi^{-1}(t) \cap \sigma^{-1}(\chi^{-1}(s))$. If $T \in \mathbf{V}$, then $\psi^{-1}(t) \in A^*\mathcal{V}$. And if $S \in \mathbf{J}_1$, then $\chi^{-1}(s)$ is a language in $B^*\mathcal{J}_1$, and hence a boolean combination of languages of the form $B^*(t, a)B^*$ ($(t, a) \in B$). Then $\sigma^{-1}(\chi^{-1}(s))$ is a boolean combination of languages of the form $\sigma^{-1}(B^*(t, a)B^*)$. Now $\sigma(a_1 \cdots a_n) \in B^*(t, a)B^*$ if and only if, for some $1 \leqslant i \leqslant n$, we have $(t, a) = (\psi(a_1 \cdots a_{i-1}), a_i)$, that is, if and only if $a_1 \cdots a_n \in \psi^{-1}(t)aA^*$. In particular, $\chi^{-1}(s)$ and $\varphi^{-1}(s, t)$ are in $A^*\mathcal{W}$, and so is any language accepted by $\varphi$. $\square$

**Remark 4.11.** The semidirect product is a powerful tool to decompose pseudovarieties. The operation $\mathbf{V} * \mathbf{W}$ is associative on pseudovarieties and Krohn and Rhodes [27] established that every finite monoid $M$ sits in an iterated product $\mathbf{X}_1 * \cdots * \mathbf{X}_k$ where each $\mathbf{X}_i$ is either $\mathbf{G}$ or $\mathbf{Ap}$ (and the $\mathbf{G}$ and $\mathbf{Ap}$ factors alternate since $\mathbf{G} * \mathbf{G} = \mathbf{G}$ and $\mathbf{Ap} * \mathbf{Ap} = \mathbf{Ap}$). This gives rise to a famous open problem, the so-called complexity problem: given $M$, can we compute the minimum number of $\mathbf{G}$ factors in a product of $\mathbf{Ap}$ and $\mathbf{G}$ containing $M$?

An analogous operation, the 2-sided semidirect product, can be used to handle the products of the form $KaL$ ( $K, L \subseteq A^*$). This time, we need to consider not only a left action of $T$ on $S$ (as for the semidirect product), but also a right action of $T$ on $S$, a map $\rho\colon S \times T \to S$, written $(s, t) \mapsto s \cdot t$, with the dual properties of a left action ($\rho_t\colon s \mapsto s \cdot t$ is an endomorphism of $S$ and $t \mapsto \rho_t$ is a morphism), and such that, for all $t, t' \in T$, $\lambda_t$ and $\rho_{t'}$ commute: $t \cdot (s \cdot t') = (t \cdot s) \cdot t'$. Then the *2-sided semidirect product* $S \mathbin{**}_{\lambda,\rho} T$ (written $S \mathbin{**} T$) is the monoid of all pairs $(s, t) \in S \times T$, with product

$$(s, t)(s', t') = (\rho(s, t')\ \lambda(t, s'),\ tt').$$

Again, the operation is extended to pseudovarieties, by letting $\mathbf{V} \mathbin{**} \mathbf{W}$ be the pseudovariety generated by the products $S \mathbin{**} T$ with $S \in \mathbf{V}$ and $T \in \mathbf{W}$. Then the following analogue of Theorem 4.10 holds.

**Theorem 4.12.** *Let $\mathcal{V}$ be a variety of languages and for each alphabet $A$, let $A^*\mathcal{W}$ is the boolean algebra generated by the languages of $A^*\mathcal{V}$ and the languages of the form $KaL$ with $K, L \in A^*\mathcal{V}$. Then the class $\mathcal{W}$ is a variety and the corresponding pseudovariety of monoids is $\boldsymbol{J}_1 \mathbin{**} \boldsymbol{V}$.*

*Proof.* The first step of the proof consists in verifying that if $K$ and $L$ are accepted by a monoid in $T \in \mathbf{V}$, then $KaL$ is accepted by $U_1^{T \times T} \mathbin{**} T$. (Note that if $K$ and $L$ are accepted by monoids $T_1$ and $T_2$, then they are both accepted by $T_1 \times T_2$, so it is no restriction to assume that $K$ and $L$ are accepted by the same monoid.) This step is performed essentially like in Lemma 4.8, and the details are left to the reader.

The second step, to prove that if $\varphi$ is a morphism $\varphi\colon A^* \to S \ast\ast T$ with $S \in \mathbf{J}_1$ and $T \in \mathbf{V}$, then each $\varphi^{-1}(s,t)$ is in $A^*\mathcal{W}$. Here too, we use (a 2-sided version of) the wreath product principle [59]. For each $a \in A$, let $\varphi(a) = (s_a, t_a)$. Let $\psi\colon A^* \to T$ be the morphism given by $\psi(a) = t_a$, let $B = T \times A \times T$ and let $\sigma\colon A^* \to B^*$ be the map

$$\sigma(a_1 \cdots a_n) = (1, a_1, \psi(a_2 \cdots a_n))\,(\psi(a_1), a_2, \psi(a_3 \cdots a_n))\,\cdots\,(\psi(a_1 \cdots a_{n-1}), a_n, 1).$$

Then, if $\chi\colon B^* \to S$ is the morphism given by $\chi(t,a,t') = (t \cdot s_a) \cdot t'$, then

$$\varphi(a_1 \cdots a_n) = (\chi\sigma(a_1 \cdots a_n),\ \psi(a_1 \cdots a_n))\,.$$

We conclude as in the proof of Theorem 4.10. $\square$

**Remark 4.13.** In view of Schützenberger's theorem (Theorem 4.2 above), one can use this result to show that the least pseudovariety closed under the operation $\mathbf{V} \mapsto \mathbf{J}_1 \ast\ast \mathbf{V}$, is the pseudovariety **Ap** of aperiodic monoids.

Semidirect product decomposition yields very difficult decision problems, such as the complexity problem briefly described in Remark 4.11. Tilson showed that the consideration of certain categories offered a systematic tool to understand semidirect (and 2-sided semidirect) product decompositions ([57], see also [49]). Almeida and Weil combined this category-theoretical approach with topological methods to provide sets of profinite identities describing many instances of semidirect products [6]. As with Mal'cev products, these sets are usually infinite and do not offer immediate solutions to decidability problems, see [7].

For the products discussed in this section, [6] gives the following descriptions.

**Proposition 4.14.** *Let **V** be a pseudovariety of monoids. Then $\mathbf{J}_1 \ast \mathbf{V}$ is defined by the set of profinite identities of the form $xy^2 = xy$ and $xyz = xzy$ for all $x, y, z \in \widehat{X^*}$ such that **V** satisfies $xy = xz = x$.*

*$\mathbf{J}_1 \ast\ast \mathbf{V}$ is defined by the set of profinite identities of the form $xy^2x' = xyx'$ and $xyzx' = xzyx'$ for all $x, y, z, x' \in \widehat{X^*}$ such that **V** satisfies $xy = xz = x$ and $yx' = zx' = x'$.*

In [6], this result is used to show the decidability of $\mathbf{J}_1 \ast \mathbf{J}$ and $\mathbf{J}_1 \ast\ast \mathbf{J}$.

It is interesting also to note that 2-sided semidirect products and category-theoretical extensions of the notion of pseudovariety can be used to decompose unambiguous products, that is, to decompose the operation $\mathbf{V} \mapsto L\mathbf{I} \circledm \mathbf{V}$, see [35].

## 5 Varieties in other algebraic frameworks

The fundamental notions explored in this chapter—classes of algebras defined by identities, properties preserved under products and quotients, *etc.*—properly belong to the domain of universal algebra. We have applied these ideas to finite monoids, ordered finite monoids, and stamps, but in fact they are applicable in a much wider variety of settings. Here we will briefly discuss some of these extensions.

The study of varieties originates in the work of Garrett Birkhoff [12], who showed that a family of algebras (defined in a very general sense) is closed under formation of subalgebras, quotients and products if and only if it is defined by a set of identities. Such families of algebras are called *varieties* because of a loose analogy with the varieties of algebraic geometry defined by sets of polynomial equations. Note that the classes of finite monoids that we have discussed are not varieties in this sense because they are not, of course, closed under infinite direct products, nor even finite quotients of infinite direct products, and consequently they cannot be defined by sets of explicit identities (as opposed to profinite identities).

Efforts to adapt Birkhoff's Theorem to finite algebras include work of Eilenberg and Schützenberger [19], and of Baldwin and Berman [8], who both showed that pseudovarieties are indeed defined by sets of identities, in the sense that an algebra belongs to a pseudovariety if and only if it satisfies all but finitely many identities of the set. A different treatment, and the one that we have followed here, based on identities in free profinite algebras, was given by Reiterman, who proved the second part of Theorem 2.15 in the setting of arbitrary finite algebras [38] (see also Banaschewski [9]).

The first part of Theorem 2.15, characterizing the language classes corresponding to pseudovarieties of finite monoids, is from Eilenberg [18]. A generalization applicable to pseudovarieties of single-sorted finite algebras is given by Almeida [3].

The ordered monoids considered in this chapter are not, strictly speaking, algebras, but rather instances of finite *$\mathcal{L}$-structures,* which are algebras together with a set of relations compatible with the operations in the algebra. Pin and Weil [36] prove an analogue of Reiterman's Theorem for such structures. In this setting the profinite identities are replaced by profinite relational identities. The profinite ordered identities discussed in this chapter are a particular instance.

Variety theories of the kind described here have also been successfully extended to a number of many-sorted algebras that arise in the domain of automata theory, and which we briefly describe:

Wilke [60] and Perrin and Pin [31] consider regular languages of infinite words. Here the corresponding algebraic objects are two-sorted algebras called *$\omega$-semigroups.* These are pairs $(S_f, S_\omega)$, where $S_f$ is a semigroup, and where there are additional operations $S_f \times S_\omega \to S_\omega$ and $S_f \to S_\omega$. Here the free object (analogous to the free monoid in the case of pseudovarieties of finite monoids) is the pair $(A^+, A^\omega)$ of finite and infinite words over $A$. The three operations correspond to ordinary concatenation of finite words, concatenation of a finite word and an infinite word to obtain an infinite word, and taking the infinite power of a finite word to obtain an infinite word.

Ésik and Weil [20, 21] describe a theory of varieties for regular languages of *ranked trees*. These are finite trees in which the nodes are labeled by letters of a finite alphabet $\Sigma$ that is the disjoint union of subalphabets $\Sigma_0, \ldots, \Sigma_n$, where the label of a node with $k$ children belongs to $\Sigma_k$. In particular, the number of children of any node in such a tree is bounded above by $n$. The corresponding algebraic objects are called *finitary preclones.* These are sequences of finite sets $S_0, S_1, \ldots$. The operation takes an element $f$ of $S_k$, and a sequence $g = (g_1, \ldots, g_k)$, where $g_i \in S_{m_i}$, and yields an element $f \cdot g$ of $S_m$, where $m = m_1 + \cdots + m_k$. The free object is the sequence $(\Sigma M_0, \Sigma M_1, \ldots)$, where $\Sigma M_k$ consists of *$k$-ary ranked trees*: these are ranked trees in which $k$ of the leaves, reading in left-to-right order, have been replaced by the variable symbol $v_1, \ldots, v_k$. In this free

preclone, the operation $f \cdot (g_1, \ldots, g_k)$ is that of replacing the $k$ variables in $f$ by the trees $g_1, \ldots, g_k$ to obtain an $m$-ary ranked tree.

The theory can be extended as well to regular languages of finite unranked forests, in which there is no bound of the degree of branching of the nodes (e.g., Bojanczyk and Walukiewicz [14], Bojanczyk, Straubing and Walukiewicz [13]). Here the corresponding algebraic objects are called *forest algebras*. These are pairs $(H, V)$ of monoids where $V$ acts on $H$. The letters $H$ and $V$ stand for 'horizontal' and 'vertical': The free object is the pair $(H_A, V_A)$ where $H_A$ consists of forests labeled by letters of $A$, and $V_A$ consists of *contexts*: forests in which the letter at one leaf has been deleted and replaced by a single variable. The product in $H_A$ is simply concatenation of forests to obtain larger forests; the product in $V_A$ is substitution of one context for the variable in another context; and the action of $V_A$ on $H_A$ is substitution of a forest for the variable in a context so as to obtain a larger forest.

For further details on this algebraic approach of the theory of regular tree languages, we refer the reader to Chapter 22 in this Handbook.

# References

[1] D. Albert, R. Baldinger, and J. Rhodes. Undecidability of the identity problem for finite semigroups. *J. Symb. Logic*, 57:179–192, 1992. 537

[2] J. Almeida. Some pseudovariety joins involving the pseudovariety of finite groups. *Semigroup Forum*, 37(1):53–57, 1988. 538

[3] J. Almeida. *Finite Semigroups and Universal Algebra*. World Scientific, Singapore, 1994. 523, 545

[4] J. Almeida and A. Azevedo. The join of the pseudovarieties of $\mathcal{R}$-trivial and $\mathcal{L}$-trivial monoids. *J. Pure Appl. Algebra*, 60(2):129–137, 1989. 538

[5] J. Almeida, A. Azevedo, and M. Zeitoun. Pseudovariety joins involving $\mathcal{J}$-trivial semigroups. *Internat. J. Algebra Comput.*, 9(1):99–112, 1999. 538

[6] J. Almeida and P. Weil. Profinite categories and semidirect products. *J. Pure Appl. Algebra*, 123(1-3):1–50, 1998. 544

[7] K. Auinger and B. Steinberg. On the extension problem for partial permutations. *Proc. Amer. Math. Soc.*, 131(9):2693–2703 (electronic), 2003. 537, 541, 544

[8] J. Baldwin and J. Berman. Varieties and finite closure conditions. *Colloq. Math.*, 35:15–20, 1976. 545

[9] B. Banaschewski. The Birkhoff theorem for varieties of finite algebras. *Algebra Universalis*, 17:360–368, 1983. 545

[10] D. A. M. Barrington, K. Compton, H. Straubing, and D. Thérien. Regular languages in $\mathrm{NC}^1$. *J. Comput. System Sci.*, 44(3):478–499, 1992. 515

[11] J. Berstel. *Transductions and context-free languages*. Teubner Studienbücher, Stuttgart, 1979. 543

[12] G. Birkhoff. On the structure of abstract algebras. *Proc. Cambridge Phil. Soc.*, 31:433–454, 1935. 545

[13] M. Bojańczyk, H. Straubing, and I. Walukiewicz. Wreath products of forest algebras, with applications to tree logics. *Logical Methods in Computer Scence*, 8:1–39, 2012. 546

[14] M. Bojańczyk and I. Walukiewicz. Forest algebras. In J. Flum and E. G. andThomas Wilke, editors, *Logic and Automata: History and Perspectives [in Honor of Wolfgang Thomas].*, volume 2 of *Texts in Logic and Games*, pages 107–132. Amsterdam University Press, 2008. 546

[15] M. J. Branco and J.-É. Pin. Equations defining the polynomial closure of a lattice of regular languages. In *Automata, languages and programming. Part II*, volume 5556 of *Lecture Notes in Comput. Sci.*, pages 115–126. Springer, Berlin, 2009. 541

[16] J. R. Büchi. Weak second-order arithmetic and finite automata. *Z. Math. Logik Grundlagen Math.*, 6:66–92, 1960. 532

[17] S. Cho and D. T. Huynh. Finite automaton aperiodicity is PSPACE-complete. *Theoret. Comp. Science*, 88:96–116, 1991. 531

[18] S. Eilenberg. *Automata, Languages, and Machines*, volume B. Academic Press, New York and London, 1976. 514, 523, 542, 545

[19] S. Eilenberg and M. Schützenberger. On pseudovarieties. *Advances in Mathematics*, 19:413–418, 1976. 545

[20] Z. Ésik and P. Weil. Algebraic recognizability of regular tree languages. *Theoret. Comp. Science*, 340:291–321, 2005. 545

[21] Z. Ésik and P. Weil. Algebraic characterization of logically defined tree languages. *Intern. J. Algebra Comp.*, 20:195–239, 2010. 545

[22] M. Gehrke, S. Grigorieff, and J.-É. Pin. Duality and equational theory of regular languages. In *Automata, languages and programming. Part II*, volume 5126 of *Lecture Notes in Comput. Sci.*, pages 246–257. Springer, Berlin, 2008. 509, 527, 528

[23] C. Glaßer and H. Schmitz. Languages of dot-depth $3/2$. *Theory Comput. Syst.*, 42(2):256–286, 2008. 531

[24] G. Higman. Ordering by divisibility in abstract algebras. *Proceedings of the London Mathematical Society. Third Series*, 2:326–336, 1952. 527

[25] J. M. Howie. *An introduction to semigroup theory*. Academic Press [Harcourt Brace Jovanovich Publishers], London, 1976. L.M.S. Monographs, No. 7. 538

[26] N. Immerman. Nondeterministic space is closed under complementation. *SIAM J. Comput.*, 17(5):935–938, 1988. 531

[27] K. Krohn and J. Rhodes. Algebraic theory of machines. I. Prime decomposition theorem for finite semigroups and machines. *Trans. Amer. Math. Soc.*, 116:450–464, 1965. 543

[28] K. Krohn, J. Rhodes, and B. Tilson. Homomorphisms and semilocal theory. In M. Arbib, editor, *The Algebraic Theory of Machines, Languages and Semigroups*. Academic Press, 1965. 541

[29] L. Libkin. *Elements of finite model theory*. Texts in Theoretical Computer Science. An EATCS Series. Springer-Verlag, Berlin, 2004. 532, 534

[30] R. McNaughton and S. Papert. *Counter-Free Automata*. The MIT Press, Cambridge, Mass., 1971. 535

[31] D. Perrin and J.-É. Pin. *Infinite words*. World Scientific, Singapore, 2004. 545

[32] J.-É. Pin. *Varieties of Formal Languages*. North Oxford Academic, London, 1986. 514, 523, 536, 540

[33] J.-É. Pin. A variety theorem without complementation. In *Russian Mathematics (Izvestija vuzov.Matematika)*, volume 39, pages 80–90, 1995. 516

[34] J.-É. Pin and H. Straubing. Some results on $\mathcal{C}$-varieties. *Theoret. Informatics Appl.*, 39:239–262, 2005. 541

[35] J.-É. Pin, H. Straubing, and D. Thérien. Locally trivial categories and unambiguous concatenation. *Journal of Pure and Applied Algebra*, 52:297–311, 1988. 544

[36] J.-É. Pin and P. Weil. A Reiterman theorem for pseudovarieties of finite first-order structures. *Algebra Universalis*, 35:577–595, 1996. 523, 541, 545

[37] J.-É. Pin and P. Weil. Polynomial closure and unambiguous product. *Theory Comput. Syst.*, 30(4):383–422, 1997. 531

[38] J. Reiterman. The Birkhoff theorem for finite algebras. *Algebra Universalis*, 14:1–10, 1982. 523, 545

[39] J. Sakarovitch. *Elements of automata theory*. Cambridge University Press, Cambridge, 2009. Translated from the 2003 French original by Reuben Thomas. 543

[40] W. J. Savitch. Relationships between nondeterministic and deterministic tape complexities. *J. Comput. System. Sci.*, 4:177–192, 1970. 530

[41] M. P. Schützenberger. On finite monoids having only trivial subgroups. *Inform. and Comput.*, 8:190–194, 1965. 538

[42] M. P. Schützenberger. Sur le produit de concaténation non ambigu. *Semigroup Forum*, 13:47–75, 1976. 540

[43] I. Simon. Piecewise testable events. In H. Barkhage, editor, *Automata Theory and Formal Languages, 2nd GI Conference, Kaiserslautern, May 22–23, 1975*, volume 33 of *LNCS*, pages 214–222. Springer, 1975. 512, 525

[44] M. Sipser. *Introduction to the Theory of Computation, 2nd Edition*. Course Technology, 2006. 530, 531

[45] B. Steinberg. On pointlike sets and joins of pseudovarieties. *Internat. J. Algebra Comput.*, 8(2):203–234, 1998. With an addendum by the author. 538

[46] B. Steinberg. On algorithmic problems for joins of pseudovarieties. *Semigroup Forum*, 62(1):1–40, 2001. 538

[47] H. Straubing. Aperiodic homomorphisms and the concatenation product of recognizable sets. *J. Pure Appl. Algebra*, 15(3):319–327, 1979. 539

[48] H. Straubing. Families of recognizable sets corresponding to certain varieties of finite monoids. *J. Pure Appl. Algebra*, 15(3):305–318, 1979. 542

[49] H. Straubing. *Finite Automata, Formal Logic, and Circuit Complexity*. Birkhäuser, Boston, Basel and Berlin, 1994. 532, 534, 535, 536, 537, 544

[50] H. Straubing, D. Thérien, and W. Thomas. Regular languages defined with generalized quantifiers. *Inform. and Comput.*, 118(2):289–301, 1995. 537

[51] R. Szelepcsényi. The method of forced enumeration for nondeterministic automata. *Acta Inform.*, 26(3):279–284, 1988. 531

[52] P. Tesson and D. Thérien. Diamonds are forever: The variety $\mathbf{DA}$. In G. M. D. Gomes Moreira Da Cunha, P. V. A. D. Silva, and J.-É. Pin, editors, *Semigroups, Algorithms, Automata and Languages, Coimbra (Portugal) 2001*, pages 475–500. World Scientific, 2002. 540

[53] D. Thérien and Thomas Wilke. Over words, two variables are as powerful as one quantifier alternation. In *STOC*, pages 234–240, 1998. 537

[54] D. Thérien and Thomas Wilke. Temporal logic and semidirect products: An effective characterization of the until hierarchy. *SIAM Journal on Computing*, 31(3):777–798, 2001. 537

[55] Thomas Wilke. Classifying discrete temporal properties. In C. Meinel and S. Tison, editors, *Proc.16th Annual Symposium on Theoretical Aspects of Computer Science (STACS'99), Trier (Germany), 1999*, number 1443 in Lecture Notes in Computer Science, pages 32–46, Heidelberg, 1999. Springer-Verlag. Invited Lecture. 536, 537

[56] Thomas Wilke. Linear temporal logic and finite semigroups. In J. Sgall, A. Pultr, and P. Kolman, editors, *Mathematical Foundations of Computer Science 2001, 26th International Symposium, MFCS 2001 Marianske Lazne, Czech Republic, August 27-31, 2001, Proceedings*, volume 2136 of *Lecture Notes in Computer Science*, pages 96–110. Springer, 2001. 537

[57] B. Tilson. Categories as algebras: An essential ingrediant in the theory of monoids. *J. Pure Appl. Algebra*, 48:83–198, 1987. 544

[58] P. G. Trotter and M. V. Volkov. The finite basis problem in the pseudovariety joins of aperiodic semigroups with groups. *Semigroup Forum*, 52(1):83–91, 1996. Dedicated to the memory of Alfred Hoblitzelle Clifford (New Orleans, LA, 1994). 538

[59] P. Weil. Closure of varieties of languages under products with counter. *J. Comput. System Sci.*, 45(3):316–339, 1992. 544

[60] T. Wilke. An algebraic theory for regular languages of finite and infinite words. *Intern. J. Algebra Comp.*, 3:447–489, 1993. 545